

АНАЛИЗ ЭФФЕКТИВНОСТИ ПРОЦЕДУР УСТОЙЧИВОГО НЕЛИНЕЙНОГО ОЦЕНИВАНИЯ ПРИ ОБРАБОТКЕ ДАННЫХ

АВРАМЕНКО В.П., ХАЛЯВИН В.А.

Исследуются основные алгоритмы робастных оценок. Выполняется сравнительный анализ робастных методов оценивания зашумленных выборок результатов наблюдений. Показывается, что за устойчивость оценок приходится расплачиваться пятипроцентной точностью; оценивание параметров в автоматизированных системах целесообразно производить совокупностью алгоритмов.

Постановка задачи исследования

Практика обработки данных в различных отраслях науки и техники показала, что снятие результатов наблюдений часто производится при значительных нарушениях исходных предпосылок научного эксперимента (нормальность, независимость, однородность), которые кладутся в основу методов математической статистики. Все это выдвигает проблему разработки методов, процедур и алгоритмов получения более загрубленных оценок параметров модели, менее чувствительных к отклонению исходных предпосылок. Такие методы оценивания параметров, нечувствительные или слабо чувствительные к структуре данных, принято называть устойчивыми или робастными [1, с. 132-137; 2, вып. 1, с. 204-217, вып. 2, с. 135-144].

Медианные оценки

Большинство устойчивых алгоритмов определения контрольного итога базируется на использовании порядковых статистик и построении взвешенных оценок. Порядковая статистика имеет место в том случае, если элементы исходного ряда наблюдений $x_1, x_2, \dots, x_n$ объема n расположены в порядке строгого возрастания $y_1 \leq y_2 \leq \dots \leq y_n$. Классической оценкой центра распределения, основанной на порядковых статистиках, является выборочная медиана $\hat{y}$, которая вычисляется по формуле

$$\hat{y} = \begin{cases} y_{k+1} & \text{при } n = 2k+1, \\ \frac{1}{2}(y_k + y_{k+1}) & \text{при } n = 2k. \end{cases}$$

Медианные оценки являются слабочувствительными к действию импульсных помех. Этот факт можно подтвердить следующим примером. Определим контрольный итог центра распределения и рассеивания ряда наблюдений 1, 2, 3, 4, 5, 6, 7, 8, 8, 10, 10 при одном “аномальном” измерении 10→53 (когда вместо последнего значения ряда 10 измерено 53). Для исходного ряда среднее арифметическое $\bar{y}_1 = 5,82$, среднеквадратическое отклонение $\sigma_{\bar{y}_1} = 0,97$, медиана $\hat{y}_1 = 6,0$, медиана абсолютных отклонений $\text{MAO}(\hat{y}_1) = 2,0$. Для измененного

ряда 1, 2, 3, 4, 5, 6, 7, 8, 8, 10, 53 имеем $\bar{y}_2 = 8,82$, $\sigma_{\bar{y}_2} = 4,6$, $\hat{y}_2 = 6,0$, $\text{MAO}(\hat{y}_2) = 2,0$. Дисперсия среднего катастрофически (в 24 раза) возросла при наличии импульсной помехи.

Слабую чувствительность медианы к “хвостам” распределений по сравнению со средним можно продемонстрировать, воспользовавшись моделью суммарной ошибки, для которой плотность распределения имеет вид [2, вып. 2, с. 137]

$$f(y) = \frac{1-p}{\sigma} \varphi\left(\frac{y}{\sigma}\right) + \frac{p}{\tau} \varphi\left(\frac{y}{\tau}\right), \quad (1)$$

причем $\tau > \sigma$, $0 \leq p \leq \frac{1}{2}$.

Плотность распределения (1) при большой вероятности $(1-p) > \frac{1}{2}$ определяется первым слагаемым – нормальной ошибкой $N(0, \sigma^2)$ с нулевым средним и дисперсией σ^2 , а при малой вероятности $(1-p) < \frac{1}{2}$ – нормальной ошибкой $N(0, \tau^2)$ с большей дисперсией τ^2 . Задав численными значениями $\sigma = 1$, $\tau = 3$, определим дисперсии среднего и выборочной медианы $\hat{y}$ по формулам [2, вып. 2, с. 137]:

$$n \text{ var}(\bar{y}) = (1-p)\sigma^2 + p\tau^2 = 1 + 9p, \quad (2)$$

$$n \text{ var}(\hat{y}) = \frac{\pi}{2} \left(1 - \frac{3}{4}p\right)^{-2}. \quad (3)$$

Задав различными значениями вероятности p , вычислим по формулам (2) и (3) дисперсии среднего и медианы и занесем их в табл. 1.

Таблица 1

p	0	0,05	0,10	0,20	0,25	0,30
Nvar $\bar{x}$	1,00	1,40	1,80	2,60	3,00	3,40
Nvar $\hat{x}$	1,57	1,69	1,83	2,14	2,36	2,60

Полученные результаты свидетельствуют о быстром ухудшении среднего и относительной устойчивости медианы. Объясняется это тем, что на значения средних в отличие от значений медианы оказывает сильное влияние даже небольшое количество “аномальных” измерений.

Винзоризованные оценки

Более устойчивыми по сравнению с медианными являются винзоризованные оценки, которые оказываются более эффективными в условиях действия импульсных помех (выбросов). Процедура оценивания включает в себя следующие шаги:

- 1) построение порядковой статистики $y_1 \leq \dots \leq y_n$;
- 2) формирование q -винзоризованной статистики путем замены первых q наблюдений на y_{q+1} , а последних q на y_{n-q} при $0 \leq q \leq \frac{1}{2}n$;
- 3) нахождение среднего, дисперсии и доверительного интервала для среднего q -винзоризованной статистики:

$$\left. \begin{aligned} z_1 &= z_2 = \dots = z_q = y_{q+1}, \\ z_{q+i} &= y_{q+i}, 2 \leq i \leq n-2q-1, \\ z_n &= z_{n-1} = z_{n-q-1} = y_{n-q} \end{aligned} \right\}$$

по следующим формулам [1, с.133]:

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i, \quad S_z^2 = \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})^2,$$

$$\bar{z} \pm t_{(1-\frac{\alpha}{2}), (h-1)} \left(\frac{n-1}{h-1} \right) \frac{S_z}{\sqrt{n}},$$

где $h = n - 1$.

Для рассмотренного ранее ряда наблюдений 1, 2, 3, 4, 5, 6, 7, 8, 8, 10, 53 обычные оценки среднего, стандартного отклонения и 95%-ного доверительного интервала для среднего соответственно равны:

$$\bar{y} = 8,82; s_{\bar{y}} = 4,6, 8,82 \pm t_{0,975;9} \frac{10}{9} * \frac{4,6}{\sqrt{11}} = (5,62; 12,02).$$

Если $q = 1$, то один-винзоризованный ряд имеет вид 2, 2, 3, 4, 5, 6, 7, 8, 8, 10, 10, а $h = n - 2q = 9$.

Соответствующие один-винзоризованные оценки среднего, стандартного отклонения и доверительного интервала равны:

$$\bar{y}_1 = 5,64, s_{\bar{y}_1} = 0,9,$$

$$5,64 \pm t_{0,975;8} \frac{10}{8} \frac{0,9}{\sqrt{11}} = (4,77; 6,51).$$

Результаты аналогичных вычислений при $q = 2, 3$ занесены в табл.2, из которой следует, что минимальная длина доверительного интервала имеет место для два-винзоризованной оценки, являющейся наиболее точной.

Таблица 2

q	$\bar{y}$	$\sigma_{\bar{y}}$	$\bar{y} \pm t_{(1-\frac{\alpha}{2}), (h-1)} \left(\frac{n-1}{h-1} \right) \frac{\sigma_{\bar{y}}}{\sqrt{n}}$	Интервал
0	8,82	4,60	(5,62; 12,02)	6,4
1	5,64	0,90	(4,77; 6,51)	1,74
2	5,73	0,69	(4,88; 6,58)	1,70
3	6,00	0,58	(4,86; 7,14)	2,28

Усеченные оценки

Усеченные оценки среднего, стандартного отклонения и доверительного интервала для среднего строятся после отбрасывания p крайних наблюдений порядковой статистики исходного ряда. Для нахождения p -усеченной оценки среднего и дисперсии воспользуемся приближенными формулами [1, с. 135]:

$$\bar{z} = \frac{1}{h} \sum_{i=p+1}^{h-p} z_i, \quad (4)$$

$$S_z^2 = \frac{1}{h-1} \sum_{i=p+1}^{h-p} (z_i - \bar{z})^2, \quad (5)$$

$$\bar{z} \pm t_{1-\frac{\alpha}{2}; h-1} \left(\frac{n-1}{h-1} \right) \frac{S_z}{\sqrt{n}}, \quad (6)$$

где $h = n - 2p$.

Для рассмотренного ранее ряда 1, 2, 3, 4, 5, 6, 7, 8, 8, 10, 53 по формулам (4)-(6) вычислим p -усеченные оценки при $p=1, 2, 3$ и определим оценку среднего с минимальным доверительным интервалом. При $p = 1$ имеем:

$$\bar{y}_1 = 5,88, S_{\bar{y}_1} = 0,92, 5,88 \pm t_{0,975;8} \left(\frac{10}{8} \right) \frac{0,92}{\sqrt{11}} = (5,08; 6,68).$$

Результаты вычислений занесены в табл. 3, из которой следует, что минимальным доверительным интервалом обладает одноусеченная оценка, она определяет контрольный итог с максимальной точностью. При выборе доли урезания обычно исходят из обеспечения надлежащей защиты от выбросов при сохранении достаточно высокой точности на выборках из нормального распределения. Практика применения урезанных оценок показала, что вполне приемлемые результаты оценивания получаются, если доля урезания выбирается из соотношения $q \approx (0,1 \div 0,2)n$.

Таблица 3

p	$\bar{z}$	$\sigma_{\bar{z}}$	$\bar{z} \pm t_{(1-\frac{\alpha}{2}), (h-1)} \left(\frac{n-1}{h-1} \right) \frac{\sigma_{\bar{z}}}{\sqrt{n}}$	Интервал
0	8,82	4,60	(5,62; 12,02)	6,40
1	5,88	0,92	(5,08; 6,68)	1,60
2	5,86	0,80	(4,88; 6,84)	1,96

Устойчивые оценки

Идея метода исходит из предположения, что задана последовательность независимых, одинаково распределенных случайных величин $x_1, \dots, x_n$ с известной априори непрерывной плотностью распределения $f(x-a)$, где a – параметр сдвига. Согласно методу максимального правдоподобия составляется мультипликативная функция правдоподобия, которая после логарифмирования записывается в виде [3, с. 12]:

$$L_n(a) = \sum_{i=1}^n \ln f(x_i - a) = - \sum_{i=1}^n \rho(x_i - a),$$

а затем отыскивается ее максимальное значение по параметру a или минимальное значение выражения

$$K(a) = \sum_{i=1}^n \rho(x_i - a), \quad \text{где } \rho(y) = -\ln f(y).$$

Поиск экстремума в предположении гладкости функции $K(a)$ находится путем дифференцирования по параметру и приравнивания полученного выражения нулю. В результате получается уравнение вида

$$\sum_{i=1}^n \psi(x_i - a) = 0 \quad \text{или в инвариантной форме}$$

$$\sum_{i=1}^n \psi\left(\frac{x_i - a}{s}\right) = 0, \quad \text{из которого определяется иско-$$

мый параметр a . Основной трудностью в получении М-оценок является выбор надлежащей функ-

ции $\psi(\cdot)$, которая обеспечивает требуемую помехоустойчивость и эффективность оценок.

Функция $\psi(\cdot)$ связана с определенным законом распределения результатов наблюдений и для некоторых из них имеет вид:

1) для нормального распределения

$$N(x, a, \sigma) = (\sigma\sqrt{2\pi})^{-1} \exp[-(x-a)^2 / 2\sigma^2]$$

$$\text{при } \sigma = 1, \quad -\ln N(x-a) \equiv \frac{1}{2}(x-a)^2 + C,$$

$$\text{при } \sigma = 1, \quad a = 0, \quad \rho(x) = \frac{1}{2}x^2 + c, \quad \psi(x) = x; \quad (7)$$

2) для двойного экспоненциального распределения

$$E(x, a, \sigma) = \frac{\sigma}{2} \exp(-\sigma|x-a|)$$

$$\text{при } \sigma = 1 \quad -\ln E(x-a) \equiv |x-a| + c,$$

$$\text{при } \sigma = 1, \quad a = 0 \quad \rho(x) = |x| + C,$$

$$\psi(x) = \begin{cases} -1 & \text{при } x < 0, \\ 1 & \text{при } x > 0; \end{cases} \quad (8)$$

3) для распределения Коши

$$K(x, a, \sigma) = \sigma / \pi(\sigma^2 + (x-a)^2)$$

$$\text{при } \sigma = 1 \quad -\ln K(x-a) = \ln(1 + (x-a)^2) + C,$$

$$\text{при } \sigma = 1, \quad a = 0 \quad \rho(x) = \ln(1 + x^2) + C,$$

$$\psi(x) = 2x / (1 + x^2). \quad (9)$$

В целях обеспечения надлежащей помехоустойчивости и эффективности оценивания функции $\psi(\cdot)$, определяемые выражениями (7), (8), (9), преобразуются определенным образом в кусочно-линейные. В практике моделирования широко используются следующие модификации этих функций [1, с.32-137].

Функции Хубера

$$\rho(x) = \begin{cases} x^2/2 & \text{при } |x| \leq 1,5, \\ 1,5|x| - \frac{1,5^2}{2} & \text{при } |x| > 1,5, \end{cases}$$

$$\psi(x) = \begin{cases} -1,5 & \text{при } x < -1,5, \\ x & \text{при } |x| \leq 1,5, \\ 1,5 & \text{при } x > 1,5 \end{cases}$$

ориентированы на обработку нормально распределенных результатов наблюдений, которым присущи “утяжеленные хвосты”, подчиненные двойному экспоненциальному закону распределения. В качестве масштаба использовано число 1,5 по той причине, что большинство элементов нормально распределенной выборки удовлетворяет условию $|x_i - a|/S < 1,5$.

Функция Андрюса

$$\psi(x) = \begin{cases} \sin(x/2,1) & \text{при } |x| \leq 2,1\pi, \\ 0 & \text{при } |x| > 2,1\pi. \end{cases}$$

Функция Хампеля

$$\psi(x) = \text{sgn}(x) * \begin{cases} |x| & \text{при } 0 \leq |x| \leq 1,7, \\ 1,7 & \text{при } 1,7 \leq |x| \leq 3,4, \\ 1,7(8,5 - |x|)/5,1 & \text{при } 3,4 \leq |x| \leq 8,5, \\ 0 & \text{при } |x| > 8,5 \end{cases}$$

приписывает крайним наблюдениям упорядоченного ряда такие веса, чтобы максимально снизить действие помех.

Устойчивые алгоритмы нахождения контрольного итога широко применяются при моделировании организационных и технологических процессов. Современное использование этих алгоритмов выполняется в сочетании с классическими процедурами. Можно воспользоваться следующей комбинированной процедурой устойчивого оценивания [4, с.25-26]:

1) оценить контрольные итоги классическим методом наименьших квадратов в предположении соблюдения всех исходных предпосылок;

2) выполнить несколько итераций оценивания с использованием функции Хубера и функции Андрюса при константе масштаба 1,5;

3) если результаты оценивания согласно п.1 и 2 совпадают, делаются статистические выводы на основании классических методов МНК; в противном случае за основу берутся “неклассические” решающие правила.

Кусочно-линейная М-оценка параметра положения (сдвига) a определяется неявно как решение уравнения

$$Q = \sum_{i=1}^n \psi_i\left(\frac{x_i - a}{S}\right) = 0, \quad (10)$$

где S — помехоустойчивая оценка параметра масштаба, в качестве которой может быть принята медиана абсолютных отклонений

$$S_1 = \text{MAO}(x) = M_B|x_i - M_B\{x_i\}|/0,6745, \quad i = \overline{1, n}$$

или интерквартильный размах

$$S_2 = \text{ИКР}(x) = (x_{0,75} - x_{0,25})/2 * 0,6745.$$

Решение уравнения (10) производится при помощи итеративной процедуры. В качестве начального приближения $a^{(0)}$ решения a можно принять медиану исходного ряда наблюдений, а оценкой параметра масштаба S может служить медиана абсолютных отклонений от $a^{(0)}$.

Определим М-оценку Хампеля для ряда наблюдений 1, 2, 3, 4, 5, 6, 7, 8, 10, 53. Выборочная медиана равна $\hat{x} = a^{(0)} = 6,0$, а медиана абсолютных отклонений — $S^{(0)} = M_B|x_i - \hat{x}| = 2,0$. После изменения масштаба по формуле

$$x_i' = (x_i - a^{(0)}) / S^0 \quad (11)$$

получим ряд -2,5; -2,0; -1,5; -1,0; -0,5; 0; 0,5; 1,0; 1,0; 2,0; 2,5, который с учетом функции Хампеля примет вид -1,7; -1,7; -1,5; -1,0; -0,5; 0; 0,5; 1,0; 1,0; 1,7.

Начальное значение невязки решения, вычисляемое по формуле (11), равно $Q^{(0)} = -2,2$. Сохраняя неизменной оценку масштаба $S^{(0)} = 2$, выбираем первое приближение оценки $a^{(1)} = 5,0$. После проведения аналогичных вычислений получим невязку после первой итерации $Q^{(1)} = 1,5$. Изменение знака невязки свидетельствует о том, что искомое решение расположено в интервале $[a^{(0)}, a^{(1)}]$. Для уточнения решения можно воспользоваться методом Ньютона, интерполяционными формулами или выполнить еще несколько итераций.

Результаты выполненных вычислений занесены в табл. 4, из которой следует, что искомое значение устойчивой М-оценки Хампеля при $Q^{(4)} = 0$ равно $a^{(4)} = 5,425$.

Таблица 4

Итерации к	0	1	2	3	4
$a^{(k)}$	6,0	5,0	5,4	5,43	5,425
$Q^{(k)}$	-2,2	1,5	-0,1	0,007	0,000

Различные оценки контрольного итога центра распределения, полученные в настоящей работе, сведены в табл. 5.

Полученные результаты позволяют сделать следующие **выводы**:

— классические методы обработки результатов наблюдений весьма чувствительны к отклонениям исходных предпосылок эксперимента;

Таблица 5

Процедура оценивания	Оценка
Классический МНК $\bar{x}$	8,82
Медианная оценка $\hat{x}$	6,00
Винзоризованная оценка	6,00
Усеченная оценка при $q = 3$	5,88
М-оценка Хампеля	5,42

— робастные методы оценивания обеспечивают высокую эффективность обработки данных (за устойчивость оценок при действии импульсных помех приходится расплачиваться пятипроцентной потерей эффективности оценок);

— оценивание параметров в автоматизированных системах целесообразно производить совокупностью алгоритмов, а отбор приемлемых решений — с помощью формализованных процедур или лицом, принимающим решение [4].

Литература: 1. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. М.: Мир, 1982. 488 с. 2. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. М.: Финансы и статистика, 1982. Вып. 1. 317с. Вып. 2. 239 с. 3. Устойчивые статистические методы оценки данных / Под ред. Р.Л. Лонера и Г.Н. Уилкоксона. М.: Машиностроение, 1984. 232 с. 4. Авраменко В.П. Устойчивые процедуры моделирования технологических процессов. АСУ и приборы автоматики. 1986. Вып. 78. С. 3-9.

Поступила в редколлегию 14.04.2000

Рецензент: д-р техн. наук, проф. Петров Э.Г.

Авраменко Валерий Павлович, д-р техн. наук, профессор кафедры информационных управляющих систем ХТУРЭ. Научные интересы: управление процессами в компьютерных системах, геоинформационные системы. Увлечения: путешествия, садоводство. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. 40-94-51.

Халявин Владимир Александрович, студент группы КСОИУ-95-1 ХТУРЭ. Научные интересы: геоинформационные системы, задачи целочисленного программирования. Увлечения: путешествия, музыка. Адрес: Украина, 61166, Харьков, ул. Бакулина, 10, ком. 429.

УДК 62.52

ОРГАНИЗАЦИЯ ПРОДУКЦИОННЫХ БАЗ ЗНАНИЙ В СИТУАЦИОННЫХ СИСТЕМАХ УПРАВЛЕНИЯ

СЫТНИК Б.Т., КАРАВАН А.Н.

Рассматривается задача организации баз знаний для диспетчерского управления движением подвижных единиц на сети городского электротранспорта. В основу ситуационной системы управления реальным временем положена модель производственной ситуации, которая на каждый момент времени отражает состояние объекта управления и окружения. Предлагаются механизмы, структуры и правила оценки производственной и модельной ситуации, а также правила модификации параметров модели.

Введение

Существует ряд методов, которые могут быть положены в основу управления перевозочным процессом на полигоне транспортной сети. Это матричные и сетевые модели, сети Петри, теория массового обслуживания и др. Каждый из этих методов имеет свои достоинства и недостатки. Так, для матричных моделей недостатком является большая размерность задачи. Модели массового обслуживания ориентированы на регулярный поток событий, следующих через равные промежутки времени, что нетипично для транспортных задач, где характерно появление нечетких последовательностей событий на участках маршрута. Поэтому решение поставленной задачи актуально в объединении методов искусственного интеллекта, сетевых моделей и ситуационного управления, позволяющих адекватно описать состояния объекта управления и окру-