УДК 681.518.54

В.В. Калачева, Д.В. Сумцов, А.М. Носик

ОЦЕНКА ЭФФЕКТИВНОСТИ ПРОЦЕДУР УСТОЙЧИВОГО НЕЛИНЕЙНОГО ОЦЕНИВАНИЯ ПРОЦЕССА ОБРАБОТКИ ДАННЫХ В СИСТЕМАХ ВООРУЖЕНИЯ

Производится оценка эффективности процедур устойчивого нелинейного оценивания процесса обработки данных в системах вооружения. Обосновывается необходимость использования робастных методов оценивания, обеспечивающих высокую эффективность обработки данных, и целесообразность произведения оценки параметров в автоматизированных системах с помощью совокупности алгоритмов.

Постановка задачи исследования

Практика обработки данных в различных отраслях науки и техники, в частности, в системах вооружения, показала, что снятие результатов наблюдений часто производится при значительных нарушениях исходных предпосылок научного эксперимента (нормальность, независимость, однородность), которые находятся в основе методов математической статистики. Все это формирует проблему разработки методов, процедур и алгоритмов получения загрубленных оценок параметров модели, которые чувствительны к отклонению исходных предпосылок. Такие методы оценивания параметров, нечувствительные или слабочувствительные к структуре данных, принято называть устойчивыми или робастными [1 – 3].

Медианные оценки

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

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

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 $\to$ 53 (когда вместо последнего значения ряда 10 измерено 53). Для исходного ряда среднее арифметическое $\overline{y}_1$ = 5,82, среднеквадратическое отклонение $\sigma_{\overline{y}_1}$ = = 0,97, медиана $\hat{y}_1$ = 6,0, медиана абсолютных отклонений (МАО) ($\hat{y}_1$) = 2,0. Для измененного ряда 1, 2, 3, 4, 5, 6, 7, 8, 8, 10, 53 имеем $\overline{y}_2$ = 8,82, $\sigma_{\overline{y}_2}$ = = 4,6, $\hat{y}_2$ = 6,0, МАО ($\hat{y}_2$) = 2,0. Дисперсия среднеквадратического катастрофически возросла (в 24 раза) при наличии импульсной помехи.

Слабую чувствительность медианы к "хвостам" распределений по сравнению со среднеквадратическим можно продемонстрировать, воспользовавшись моделью суммарной ошибки, для которой плотность распределения имеет вид [3]

$$f(y) = \frac{1-p}{\sigma}\varphi\left(\frac{y}{\sigma}\right) + \frac{p}{\tau}\varphi\left(\frac{y}{\tau}\right), \quad (2)$$

причем $\tau > \sigma$, $0 \le p \le 1/2$.

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

$$n\,\mathrm{var}(\overline{y}) = (1-p)\sigma^2 + p\tau^2 = 1 + \sigma p\,; \quad (3)$$

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

Задавшись различными значениями вероятности p, вычислим по формулам (3) и (4) дисперсии среднеквадратического и медианы и занесем их в табл. 1.

Полученные результаты свидетельствуют о быстром ухудшении дисперсии среднеквадратического

Таблица 1

Дисперсии среднеквадратического и медианы

p	0	0,01	0,05	0,10	0,15	0,20	0,25	0,30
nvar $\overline{x}$	1,00	1,08	1,40	1,80	2,20	2,60	3,00	3,40
nvar €	1,57	1,59	1,69	1,83	1,98	2,14	2,36	2,60

Таблица 2

Результаты вычислений при q = 0, 1, 2, 3

q	$\overline{y}$	$\sigma_{\overline{y}}$	$\overline{y} \pm t_{\left(1-\frac{\alpha}{2}\right),(h-1)} \left(\frac{n-1}{h-1}\right) \frac{\sigma_{\overline{y}}}{\sqrt{n}}$,	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

и относительной устойчивости медианы. Объясняется это тем, что на значения среднеквадратических 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 $y_1 \le \ldots \le y_n$;

2) формирование q-винзоризованной статистики путем замены первых q наблюдений на y_{q+1}, а последних q на y_{n-q} при $0 \le q \le \frac{1}{2} n$;

3) нахождение среднеквадратического, дисперсии и доверительного интервала для среднеквадратического q-винзоризованной статистики

$$\left.\begin{aligned} & z_1 = z_2 = \ldots = z_q = y_{q+1}; \\ & z_{q+i} = y_{q+i},\ 2 \le i \le n-2q-1; \\ & z_n = z_{n-1} = z_{n-q-1} = y_{n-q} \end{aligned}\right\} \qquad (5)$$

по следующим формулам [2]:

$$\overline{z} = \frac{1}{n} \sum_{i=1}^{n} z_i\ ; \qquad (6)$$

$$S_{\overline{z}}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \overline{z})^2; \qquad (7)$$

$$\overline{z} \pm t_{\left(1-\frac{\alpha}{2}\right),(h-1)} \left(\frac{n-1}{h-1}\right) \frac{S\overline{z}}{\sqrt{n}}, \qquad (8)$$

где $h = n - 1$.

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

$$\overline{y} = 8{,}82;\ s_{\overline{y}} = 4{,}6;$$

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

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 среднеквадратического, стандартного отклонения и доверительного интервала равны:

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

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

Результаты аналогичных вычислений при q = 0, 1, 2, 3 занесены в табл. 2, из которой следует, что минимальная длина доверительного интервала имеет место для два-винзоризованной оценки, являющейся наиболее точной.

Усеченные оценки

Усеченные оценки среднеквадратического, стандартного отклонения и доверительного интервала для среднеквадратического строятся после отбрасывания p крайних наблюдений порядковой стати-

Таблица 3

Результаты вычислений при p = 1, 2

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

стики исходного ряда. Для нахождения p-усеченной оценки среднеквадратического и дисперсии воспользуемся приближенными формулами [1]:

$$\bar{z} = \frac{1}{h}\sum_{i=p+1}^{h-p} z_i; \qquad (11)$$

$$S_{\bar{z}}^2 = \frac{1}{h-1}\sum_{i=p+1}^{h-p}(z_i - \bar{z})^2; \qquad (12)$$

$$\bar{z} \pm t_{1-\frac{\alpha}{2};h-1}\left(\frac{n-1}{h-1}\right)\frac{S_{\bar{z}}}{\sqrt{n}}, \qquad (13)$$

где $h = n - 2p$.

Для рассмотренного ранее ряда 1, 2, 3, 4, 5, 6, 7, 8, 8, 10, 53 по формулам (11) – (13) вычислим p-усеченные оценки при p = 1, 2 и определим оценку среднего с минимальным доверительным интервалом. При p = 1 имеем

$$\bar{y}_1 = 5{,}88; \; S_{\bar{y}_1} = 0{,}92;$$

$$5{,}88 \pm t_{0,975;8}\left(\frac{10}{8}\right)\frac{0{,}92}{\sqrt{11}} = (5{,}08; 6{,}68). \qquad (14)$$

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 $q \approx (0{,}1 \div 0{,}2)n$.

Устойчивое оценивание на основе кусочно-линейных М-оценок

Идея метода исходит из предположения, что задана последовательность независимых, одинаково распределенных случайных величин $x_1, \ldots, x_n$ 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]

$$L_n(a) = \sum_{i=1}^{n} \ln f(x_i - a) = -\sum_{i=1}^{n}\rho(x_i - a), \qquad (15)$$

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

$$K(a) = \sum_{i=1}^{n}\rho(x_i - a), \qquad (16)$$

где

$$\rho(y) = -\ln f(x). \qquad (17)$$

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

$$\sum_{i=1}^{n}\psi(x_i - a) = 0 \qquad (18)$$

или в инвариантной форме

$$\sum_{i=1}^{n}\psi\left(\frac{x_i - a}{s}\right) = 0, \qquad 19)$$

из которого определяется искомый параметр a.

Основной трудностью в получении М-оценок является выбор надлежащей функции $\psi(\cdot)$, которая обеспечивает требуемую помехозащищенность и эффективность оценок.

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

1) для нормального распределения

$$N(x,a,\sigma) = (\sigma\sqrt{2\pi})^{-1}\exp[-(x-a)^2/2\sigma^2]; \quad (20)$$

$$-\ln N(x-a) \equiv \frac{1}{2}(x-a)^2 + C; \qquad (21)$$

$$\rho(x) = \frac{1}{2}x^2 + c; \quad \psi(x) = x \qquad (22)$$

при σ = 1, a = 0;

2) для двойного экспоненциального распределения

$$E(x,a,\sigma) = \frac{\sigma}{2}\exp(-\sigma|x-a|) ; \quad (23)$$

$$-\ln E(x-a) \equiv |x-a| + c; \quad (24)$$

$$\rho(x) = |x| + C; \quad \psi(x) = \begin{cases} -1 \text{ при } x < 0; \\ 1 \text{ при } x > 0 \end{cases} \quad (25)$$

при $\sigma = 1$, $a = 0$;

3) для распределения Коши

$$K(x,a,\sigma) = \sigma / \pi(\sigma^2 + (x-a)^2) ; \quad (26)$$

$$-\ln K(x-a) = \ln(1+(x-a)^2) + C; \quad (27)$$

$$\rho(x) = \ln(1+x^2) + C; \quad \psi(x) = 2x/(1+x^2) \quad (28)$$

при $\sigma = 1$, $a = 0$.

С целью обеспечения надлежащей помехоустойчивости и эффективности оценивания функции $\psi(\cdot)$, описываемые выражениями (22), (25), (28), преобразуются определенным образом в кусочно-линейные. В практике моделирования широко используются следующие модификации этих функций [1].

Функции Хубера

$$\rho(x) = \begin{cases} x^2/2 & \text{при } |x| \le 1{,}5; \\ 1{,}5|x| - \dfrac{1{,}5^2}{2} & \text{при } |x| > 1{,}5, \end{cases} \quad (29)$$

$$\psi(x) = \begin{cases} -1{,}5 \text{ при } x < -1{,}5; \\ x \quad \text{при } |x| \le 1{,}5; \\ 1{,}5 \quad \text{при } x > 1{,}5 \end{cases} \quad (30)$$

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 удовлетворяют условию $|x_i - a|/S < 1{,}5$.

Функция Андрюса

$$\psi(x) = \begin{cases} \sin(x/2{,}1) & \text{при } |x| \le 2{,}1\pi; \\ 0 & \text{при } |x| > 2{,}1\pi. \end{cases} \quad (31)$$

Функция Хампеля

$$\psi(x) = $$

$$= \operatorname{sgn}(x)\begin{cases} |x| & \text{при } 0 \le |x| \le 1{,}7; \\ 1{,}7 & \text{при } 1{,}7 \le |x| \le 3{,}4; \\ 1{,}7(8{,}5-|x|)/5{,}1 & \text{при } 3{,}4 \le |x| \le 8{,}5; \\ 0 & \text{при } |x| > 8{,}5 \end{cases} \quad (32)$$

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 [3]:

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 п.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

$$Q = \sum_{i=1}^{n} \psi_i\left(\frac{x_i - a}{S}\right) = 0, \quad (33)$$

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

$$S_1 = \text{МАО}(x) = M_в|x_i - M_в\{x_i\}| / 0{,}6745, \; i = \overline{1,n} \quad (34)$$

или интерквартельный размах

$$S_2 = \text{ИКР}(x) = (x_{0,75} - x_{0,25}) / 2 \cdot 0{,}6745. \quad (35)$$

Решение уравнения (33) 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, 8, 10, 53. Выборочная медиана равна $\hat{x} = a^{(0)} = 6{,}0$, а медиана абсолютных отклонений – $S^{(0)} = M_в|x_i - \hat{x}|$ = 2,0. После изменения масштаба по формуле

$$x_i^{'} = (x_i - a^{(0)}) / S^0 \quad (36)$$

получим ряд –2,5; –2,0; –1,5; –1,0; –0,5; 0; 0,5; 1,0; 1,0; 2,0; 23,5, который с учетом функции Хампеля (32) 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 (33), равно $Q^{(0)}$ = –2,2. Со-

Таблица 4

Результаты выполненных вычислений М-оценки Хампеля

Итерации к	0	1	2	3	4
$a^{(к)}$	6,000	5,000	5,400	5,430	5,425
$Q^{(к)}$	–2,200	1,500	–0,100	0,007	0,000

Таблица 5

Различные оценки контрольного итога центра распределения

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

храняя 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, 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 $a^{(4)} = 5{,}425$. Различные оценки контрольного итога центра распределения, полученные в настоящей работе, сведены в табл. 5.

Выводы

На основе полученных данных можно сделать следующие выводы:

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

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

оценивание параметров в системах вооружения целесообразно производить совокупностью алгоритмов, а отбор приемлемых решений – с помощью формализованных процедур или лицом, принимающим решение [5].

СПИСОК ЛИТЕРАТУРЫ

1. Авраменко В.П., Виноградов Г.М., Демченко Е.О. Моделирование управления городским электротранспортом // Інформаційно-керуючі системи на залізничному транспорті: Науково-технічний журнал. – 1997. – № 1. – С. 66 – 69.

2. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. – М.: Мир, 1982. – 488 с.

3. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. – М.: Финансы и статистика, 1982. – Вып. 1. – 317 с. – Вып. 2. – 239 с.

4. Устойчивые статистические методы оценки данных / Под ред. Р.Л. Лонера и Г.Н. Уилкоксона. – М.: Машиностроение, 1984. – 232 с.

5. Авраменко В.П. Устойчивые процедуры моделирования технологических процессов // АСУ и приборы автоматики. – 1986. – Вып. 78. – С. 3 – 9.

Поступила 14.12.2005

Рецензент: д-р физ.-мат. наук профессор С.В. Смеляков, Харьковский университет Воздушных Сил.