УДК 519.281.1

В. А. ЕГОРОВ, Ю. Г. МОРОЗОВ
(Ленинград)

АЛГОРИТМ УДАЛЕНИЯ АНОМАЛЬНЫХ ТОЧЕК В ЭКСПЕРИМЕНТАЛЬНОМ МАТЕРИАЛЕ ПРИ АВТОМАТИЗИРОВАННОЙ РЕГИСТРАЦИИ ДАННЫХ

При использовании в научном эксперименте системы автоматизированного сбора и обработки данных с использованием ЭВМ возникает проблема ликвидации ошибок, возникающих вследствие случайных внешних воздействий на экспериментальную установку и регистрирующую аппаратуру. Хотя обычно число аномальных точек в экспериментальном материале невелико, они могут внести существенные искажения в результат дальнейшей обработки, так как отклонения этих точек от истинного значения могут быть очень большими.

В данной работе предлагается алгоритм удаления аномальных точек, ориентированный на системы, в которых обработка проводится после накопления большой совокупности данных (системы типа «off line», регистрация быстрых процессов с помощью систем типа «on line» и др.).

1. **Постановка задачи.** Пусть в процессе измерений регистрируется некоторая функция $Y(Z)$ в виде двух точечных множеств $\{Y_i\}_{i=1}^n$, $\{Z_i\}_{i=1}^n$ где i — номер измерения, соответствующий моменту времени t_i. Представим наблюдаемую экспериментальную функцию в следующем виде:

$$Y(t) = f_1(t) + x(t),$$
$$Z(t) = f_2(t).$$

Здесь f_1 и f_2 — истинные значения соответствующих функций, причем функция $f_2(t)$ монотонна, а $x(t)$ — ошибка измерения.

Будем предполагать, что измерения проводятся с помощью «прибора», который с вероятностью δ дает «сбой», т. е. функция распределения ошибки измерения имеет вид

$$F(x) = (1 - \delta)G(x) + \delta H(x),$$

причем случайная величина с функцией распределения $H(x)$ принимает «большие» значения, и ее разброс много больше разброса величины с функцией распределения $G(x)$. Экспериментальные точки, соответствующие сбоям прибора, будем называть аномальными. При наличии аномальных точек непосредственное использование стандартных способов обработки дает плохие результаты (метод наименьших квадратов и т. п.). Поэтому в случае, когда целью задачи является оценка параметров, от которых зависит функция f, есть смысл использовать вычислительный алгоритм, состоящий из двух этапов: на первом проводится отбраковка наблюдений, позволяющая избавиться от аномальных точек, а на втором — «стандартная» обработка данных.

Описание алгоритма удаления аномальных точек. Пусть имеется экспериментальная зависимость $Y(Z)$, заданная в виде двух числовых последовательностей

$$\{Z_i\}_{i=1}^n, \ \{Y_i\}_{i=1}^n,$$

где i — номер экспериментальной точки. Пусть далее известно, что в процессе измерения с вероятностью δ на кривой появляются резкие выбросы (аномальные точки).

Предлагаемая последовательность действий по отбраковке аномальных точек заключается в следующем: из множества номеров точек слу-

чайным образом независимо выбирается m подмножеств по k элементов:

$$\{i_p^{(j)}\}_{p=1}^{k}, \quad j = 1, 2, \ldots, m.$$

Для каждого из подмножеств строится приближение функции $f(z)$ следующим образом:

$$\hat{f}^{(j)}(Z_{i_p}^{(j)}) = Y_{i_p}^{(j)}, \quad p = 1, 2, \ldots, k.$$

По каждой паре соседних $\hat{f}^{(j)}(Z_{i_p}^{(j)})$ с помощью линейной интерполяции вычисляются приближения в остальных точках $\hat{f}^{(j)}(Z_i)$. Предполагается, что полученная для каждой выборки ломаная линия при отсутствии в этой выборке аномальных точек повторяет истинный ход экспериментальной функции с точностью до некоторого ε. Далее, тем точкам, для которых выполняется неравенство

$$|Y_i - \hat{f}^{(j)}(Z_i)| > \varepsilon,$$

приписывается штрафное очко. Все указанные операции проделываются для $j = 1, 2, \ldots, m$. Затем точки, сумма накопленных штрафов у которых превосходит некоторое пороговое число l_{0i}, исключаются из совокупности экспериментальных данных.

Некоторые количественные оценки параметров алгоритма. Параметрами алгоритма являются n — число экспериментальных точек, k — число членов одной независимой выборки, m — число проведенных испытаний, ε — требуемая точность совпадения Y_i с $\hat{f}^{(j)}(Z_i)$, l_{0i} — пороговое число штрафов для i-й точки.

Пусть i — номер аномальной точки. Практически эта точка штрафуется всегда, за исключением тех случаев, когда (Y_i, Z_i) является одной из вершин аппроксимирующей ломаной или находится за пределами выборки. Вероятность того, что точка попадает в случайную выборку, равна k/n. Обозначим

$$Z_{\min}^{(j)} = \min_{p \leqslant k} Z_{i_p}^{(j)}, \quad Z_{\max}^{(j)} = \max_{p \leqslant k} Z_{i_p}^{(j)},$$

где $(i_1^{(j)}, \ldots, i_k^{(j)})$ —элементы j-й выборки, а $k < n/2$. Тогда вероятность того, что i-я точка окажется вне интервала $(Z_{\min}^{(j)}, Z_{\max}^{(j)})$, равна

$$\alpha_i = \begin{cases} \dfrac{C_{n-i}^k}{C_n^k}, & i \leqslant k, \\ \dfrac{C_{i-1}^k + C_{n-i}^k}{C_n^k}, & k < i \leqslant n-k, \\ \dfrac{C_{i-1}^k}{C_n^k}, & i > n-k. \end{cases}$$

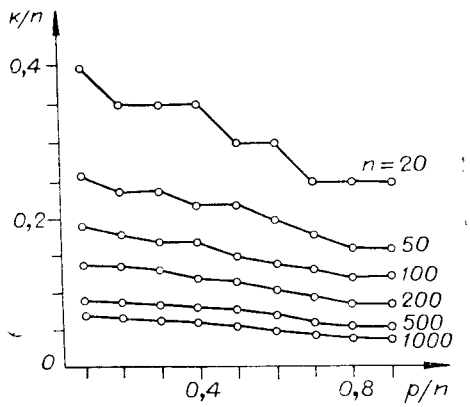
В таком случае количество штрафных очков, полученных аномальной точкой с номером i за m розыгрышей, имеет биномиальное распределение

$$P(l_i = j) = C_m^j \beta_i^j (1 - \beta_i)^{m-j},$$

где $\beta_i = 1 - k/n - \alpha_i$, а l_i — число штрафов, накопленное i-й точкой. Следовательно, среднее число штрафов, получаемое аномальной точкой с номером i, определяется соотношением

$$a_{0i} = El_i = m\beta_i. \tag{1}$$

Рассмотрим теперь результаты применения алгоритма к точкам, не являющимся аномальными. Пусть i — номер «хорошей» точки. Если не учитывать погрешность аппроксимации, то хорошая точка может быть



Зависимость оптимального значения k, отнесенного к n, от величины p/n для различных значений n.

оштрафована только в том случае, если одна из ближайших к ней вершин ломаной является аномальной точкой. Вероятность последнего события равна с точностью до $O(\delta^2)$:

$$\Theta_i \simeq 2\delta(1-\alpha_i).$$

Обозначим через λ вероятность того, что хорошая точка получит штраф вследствие ошибки аппроксимации. Предположим дополнительно, что теоретическая зависимость $Y(Z)$, представленная в виде точечного множества $(\overline{Y}_q, Z_q)$, во всей области допустимых значений аппроксимируется отрезком прямой с погрешностью не больше ε на любом интервале $[Z_q, Z_{q+\tau}]$, где $\tau = 1, 2, \ldots, p$; тогда

$$\lambda \leqslant \frac{C_{n-p}^k}{C_n^k}.$$

Далее для упрощения последующих рассуждений будем полагать

$$\lambda \approx \frac{C_{n-p}^k}{C_n^k},$$

т. е. будем считать, что хорошая точка штрафуется всегда, когда интервал, по которому происходит линейная интерполяция, больше интервала $[Z_q, Z_{q+p}]$. Полная вероятность получения штрафа хорошей точкой будет равна

$$\chi_i = 2\delta(1-\alpha_i) + \lambda.$$

Как и раньше, число штрафов имеет биномиальное распределение и среднее число штрафов, полученное хорошей точкой, равно

$$a_{1i} = m\chi_i \approx 2\delta m(1-\alpha_i) + \lambda m. \qquad (2)$$

При выборе значения параметра l_{0i} важно использовать метод Неймана — Пирсона для сравнения гипотез о распределении штрафов (биномиальное с параметром β_i или биномиальное с параметром χ_i).

Для оценки параметра k примем в качестве критерия максимум выражения

$$A = \sum_{i=1}^{n} (a_{0i} - a_{1i}). \qquad (3)$$

Подставляя (1) и (2) в (3) и учитывая, что

$$\sum_{i=1}^{n} \alpha_i = \frac{2(n-k)}{k+1},$$

получим

$$A = m\left[(1-2\delta)\,n - k - (1-2\delta)\frac{2(n-k)}{k+1} - n\frac{C_{n-p}^k}{C_n^k}\right]. \qquad (4)$$

Численный анализ выражения (4) показал, что значение k, при котором достигается максимум величины A, слабо зависит от параметра δ при изменении последнего в реально приемлемых границах [0,01, 02]; следует

учесть при этом, что мы обычно не имеем точных сведений о величине δ. На рис. 1 изображена совокупность точек, представляющая результат численного анализа (4) при $\delta = 0{,}1$. Изображенное на рисунке семейство кривых вполне пригодно для выбора оптимального значения k, для $n \in [20, 1000]$ и для $\delta \in [0{,}01, 0{,}2]$.

В том случае если в выражении (4) можно пренебречь последним членом, то близкое к оптимальному значение k определяется соотношением

$$k = [-1 + \sqrt{2(n+1)(1-2\delta)}].$$

Параметр ε в общем случае следует выбирать из условия

$$\varepsilon > 2\sigma,$$

где σ^2 — дисперсия ошибки измерений при отсутствии сбоев.

В заключение отметим, что вышеописанный алгоритм отбраковки аномальных точек весьма просто реализуется на ЭВМ и в настоящее время имеются различные варианты программ.

Поступила в редакцию 22 сентября 1978 г.