ISSN 0868–5886 *НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2009, том 19, № 3, с. 35–40*

ТЕОРЕТИЧЕСКИЕ ИССЛЕДОВАНИЯ

УДК 621.391.26

© В. В. Манойлов, И. В. Заруцкий

ИССЛЕДОВАНИЕ АЛГОРИТМОВ ОТБРАКОВКИ ВЫБРОСОВ В МАСС-СПЕКТРОМЕТРИЧЕСКИХ СИГНАЛАХ

Описываются результаты исследования нескольких типов алгоритмов отбраковки выбросов и сглаживания в масс-спектрометрических сигналах. Исследование проведено с помощью вычислительного эксперимента с использованием компьютерных моделей масс-спектрометрических сигналов, искаженных шумами и выбросами. Показаны преимущества модифицированного для масс-спектрометрии алгоритма Rousseeuw, основанного на вычислении квадратов медиан разности текущих отсчетов сигналов.

Кл. сл.: методы обработки сигналов, масс-спектрометрия, статистический анализ данных, сглаживание и фильтрация сигналов

ПОСТАНОВКА ЗАДАЧИ

Одной из важных задач обработки сигналов масс-спектрометрического анализа является удаление выбросов, которое имеет существенное значение для повышения точности измерений. Мы наблюдаем мультимодальную функцию (спектр) в аддитивной смеси со стационарным шумом и выбросами (ложными наблюдениями). Требуется освободиться от выбросов и оценить дисперсию (среднеквадратическое отклонение) шума при минимально возможном искажении мультимодальной функции.

Под "выбросами", или "ложными элементами", понимают данные, сильно отличающиеся по величине от математического ожидания анализируемой выборки случайной величины. Плотность распределения данных случайных величин, в которых имеются выбросы f_v обычно представляется в виде распределения Тьюки, которое записывается в следующем виде:

$$f_v = f_1(1-\alpha) + f_2\alpha ,$$

где f_1 — плотность распределения данных, в которых отсутствуют выбросы; f_2 — плотность распределения данных, которые являются выбросами, α — относительное количество выбросов в анализируемой выборке данных. Например, если в анализируемой выборке данных содержится 10 % выбросов, то $\alpha = 0.1$, если 5 %, то $\alpha = 0.05$. В случае нормального распределения анализируемых данных

$$f_v = N(m_1,\sigma_1)(1-\alpha) + N(m_2,\sigma_2)\alpha ,$$

где f_1 и f_2 представляют собой функции Гаусса со средними значениями соответственно m_1, m_2 и средними квадратичными отклонениями σ_1, σ_2. Очень часто выбросами являются данные, плотность распределения которых имеет среднее значение $m_2 = m_1$, а средние квадратичные отклонения отличаются в k раз, т. е. $\sigma_2 = k\sigma_1$. При этом число k может быть существенно больше 3. Для отбраковки выбросов наиболее простым и достаточно надежным методом является метод "3σ" (трех сигм). При наличии в выборке данных выбросов оценка значения величины σ может быть искажена. Описываемые ниже алгоритмы показывают, каким образом в задачах масс-спектрометрического анализа производится оценка значения величины σ.

В работе Rousseeuw P.J. [1] дается изящный и простой метод отыскания σ, основанный на определении $\min \mathrm{med}(r_i^2)$, где r_i^2 — квадрат разности экспериментальных наблюдений в выборке заданного размера N: $[y_1 ... y_i ... y_N.]$, называемой скользящим окном; $M = \min \mathrm{med}(r_i^2)$ является статистикой, для которой наблюдение $y_i = \hat{a} + M$ является верным (не выбросом); $\hat{a}$ — оценка функции (в работе [1] константы), на которую наложены шум и выбросы. Для нахождения $\min \mathrm{med}(r_i^2)$ необходимо:

– построить вариационный ряд, т. е. упорядочить наблюдения по величине

$$y_1 < y_2 < \ldots < y_N ;$$

– составить разности

$$\Delta h = y_{[N/2]+h} - y_h, \quad h = 1, 2, \ldots, [N/2];$$

– найти среди них минимальную $\Delta h0$

и тогда соответствующая этой разности полусумма $(y_{[N/2]+h0} + y_{h0})/2$ будет являться оценкой функции $\hat{a}$, а разность $(y_{[N/2]+h0} - y_{h0})$ является оценкой статистики M, которая реализует $\min \operatorname{med}(r_i^2)$.

Все же основная идея метода минимума квадрата медиального отклонения для нашей задачи должна быть в определенной степени доработана. Основная идея доработки метода минимума квадрата медианного отклонения для масс-спектрометрии принадлежит Сирвидасу С.И. [2]. Суть доработки состоит в том, что масс-спектр является быстроизменяющейся функцией, следовательно, нужна оценка текущего значения первой и второй производной и, кроме того, требуется дополнение в виде процедуры типа скользящего окна, т. е. внедрение локального сглаживания [2] и определение оптимальных размеров скользящего окна по критерию минимальной суммарной дисперсии. Ниже описывается доработанный метод, представленный в виде алгоритма, и результаты исследования его возможностей в сравнении с другими алгоритмами, решающими задачу отбраковки выбросов.

АЛГОРИТМ МЕТОДА МИНИМУМА КВАДРАТА МЕДИАННОГО ОТКЛОНЕНИЯ, МОДИФИЦИРОВАННЫЙ ДЛЯ ОБРАБОТКИ МАСС-СПЕКТРОМЕТРИЧЕСКИХ СИГНАЛОВ

1. Ввод данных.
2. Формируем массив x_i, $i = [1, n]$.
3. Цикл i от 1 до n.
4. Цикл w (ширина окна) от 3 до $w_{\max}$ (w — нечетное).
5. $y_j = x[i + j], \quad j = \overline{1, w}$.
6. Вычисляем $\mathrm{d}y_j = \dfrac{y_{j+2} - y_j}{2}$.
7. Строим вариационный ряд $\mathrm{d}\tilde{y}_j$.
8. Вычисление $y^1 = \operatorname{med} \mathrm{d}\tilde{y}_j$.
9. Вычисляем $\mathrm{d}^2 y_j = \dfrac{y_{j+2} - 2y_{j+1} + y_j}{2}$.
10. Строим вариационный ряд $\mathrm{d}^2 \hat{y}_j$.
11. Вычисляем $y^2 = \operatorname{med} \mathrm{d}^2 \hat{y}_j$.
12. Вычисляем $z_j = y_j - y^1(j - h) - y^2 \dfrac{(j-h)^2}{2}$, где $h = \left[\dfrac{w}{2}\right]$.
13. Строим вариационный ряд $\tilde{z}_j$.
14. Вычисляем j_0 такое, что $\tilde{z}[j_0 + h + 1] = \min(\tilde{z}[j + h + 1] - \tilde{z}[j]); \quad j = [1, h]$.
15. Вычисляем $\alpha_w = \dfrac{\tilde{z}[j_0 + h + 1] + \tilde{z}[j_0]}{2}$.
16. Вычисляем $\sigma_w = \sqrt{\dfrac{(\tilde{z}[j_0 + h + 1] - \tilde{z}[j_0])^2}{2}} \cdot 1.483$.
17. Проверим условие $\dfrac{|z_j - \alpha_w|}{\sigma_w} \le 3$.
18. Если п. 17 выполнен, заносим индекс j в массив I.
19. Вычисляем S — число элементов в массиве I.
20. Вычисляем $b_w = \dfrac{1}{S} \sum_{j \in \mathrm{I}} z_j$.
21. Вычисляем $r_w = \sum_{j \in \mathrm{I}} (z_j - b_w)^2$.
22. Для всех $j \notin \mathrm{I}$ и $1 \le i + j \le N$ увеличиваем элемент массива J[$i + j$] на 1.
23. Повторим шаги 5–22 для следующих w.
24. Определяем w_0 такое, что $r_{w_0} = \min\limits_w r_w$.
25. Значение b_w, соответствующее $r_{w\min}$, заносим в новый массив "чистых" данных $\tilde{x}_i = b_{w_0}$.
26. Повторяем шаги 4–26 для следующих I.

Поясним некоторые шаги алгоритма.

Шаг 12. Для работы данного метода в масс-спектрометрии необходимо оценить значение функции в данной точке. Имея значения первой и второй производных в каждой точке, можно воспользоваться разложением Тейлора

$$y(x) \cong y(x_0) + y'(x_0)\frac{x}{1!} + y''(x_0)\frac{x^2}{2!} \ldots,$$

откуда значение $y(x_0)$ есть разность

$$y(x_0) \cong y(x) - y'(x_0)\frac{x}{1!} - y''(x_0)\frac{x^2}{2!}.$$

Таким образом, получаем оценку функции в окне размером w ($w = 2k + 1$ для $k = 0, \pm1, \pm2, \pm3 \ldots \pm k$):

$$\hat{a} = z(S) = y_{N+S} - y'_{(S)} S - y''_{(S)} \frac{S^2}{2} \approx \text{const}.$$

Шаг 16. Оценка среднего квадратичного значения σ основывается на том факте, что интервал $[\tilde{a} - M, \tilde{a} + M]$ содержит половину наблюдений y_i ("верные" наблюдения) и соответственно интервал $[-M, M]$ содержит половину остатков r_i. Из теории вероятности известно, что радиус интервала, в который попадает половина значений случайной величины — вероятное отклонение

(В.О.) от математического ожидания нормальной величины с дисперсией σ^2 равен 0.6745σ, т. к. значение функции ошибок $\mathrm{erf}(0.6745/\sqrt{2}) = 0.5$, что означает, что вероятность

$$P(|x| < 0.6745\sigma) = 0.5 \text{ при } \sigma = 1.$$

Отсюда оценка σ: $\hat{\sigma} = \dfrac{M}{0.6745} = 1.483M$.

Адаптивность алгоритма заключается в том, что из всех полученных значений b_w при различной ширине окна W в качестве оценки берется значение с индексом W_0, которое доставляет $\min\limits_{w} \sum\limits_{1}^{w} (y_i - b_w)^2$.

После окончания работы алгоритма в массиве J для каждого элемента массива исходных данных хранится количество его упоминаний в качестве выброса *i*.

Ниже описываются результаты исследования доработанного для масс-спектрометрических сигналов метод отбраковки выбросов.

МОДЕЛИРОВАНИЕ "ЗАГРЯЗНЕННОГО" СИГНАЛА И ОБРАБОТКА ДАННОГО СИГНАЛА С ПРИМЕНЕНИЕМ РАЗЛИЧНЫХ МЕТОДОВ ОТБРАКОВКИ ВЫБРОСОВ

В основе модели лежит гауссова функция вида

$$y = \sum_{j=-200}^{200} \exp\left(-\frac{(t - t_j)^2}{2\mu^2}\right). \qquad (1)$$

Параметры функции вводятся в программу с последующей записью результатов. Далее моделируется случайный шум с нормальным распределением с заданным СКО. В основе модели создания шума лежит функция формирования случайного числа. Суммируются 12 случайных чисел, из суммы вычитается 6, полученное число умножается на СКО — результатом является шум, который добавляется к значениям функции y_i. Далее увеличиваем *i* на 1 и повторяем процедуру.

В основе модели создания выбросов лежит функция формирования случайного числа. Из датчика случайных чисел берется число и, если оно меньше, чем процент выбросов, то к "зашумленной" уже функции при текущем значении *i* прибавляет выброс с заданной амплитудой. После этого производится "очистка" сигнала при помощи метода параметрического сглаживания. Параметры модели гауссовой функции:

$$A = 30;\ \mu = 3;\ t = 0.$$

Для проведения математического эксперимента на языке программирования C++ написана программа. В качестве функций для тестирования выбраны следующие три варианта:

1) исходная гауссова функция (параметры указаны выше);

2) исходная функция с наложенным шумом и

3) с выбросами.

СКО шума — 0.0005, % выбросов (% выбр.) и их амплитуда (Ампл.) вводились в программу в диалоговом режиме.

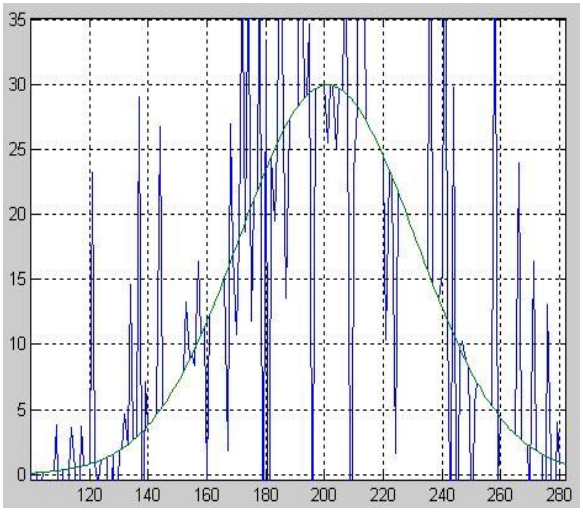

Рис. 1. Исходный сигнал с ложными выбросами

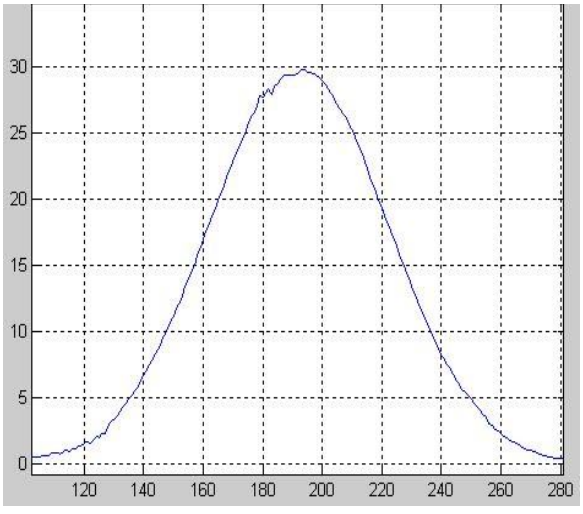

Рис. 2. Сигнал после отбраковки ложных выбросов алгоритмом минимизации квадрата медианного остатка (МКМО)

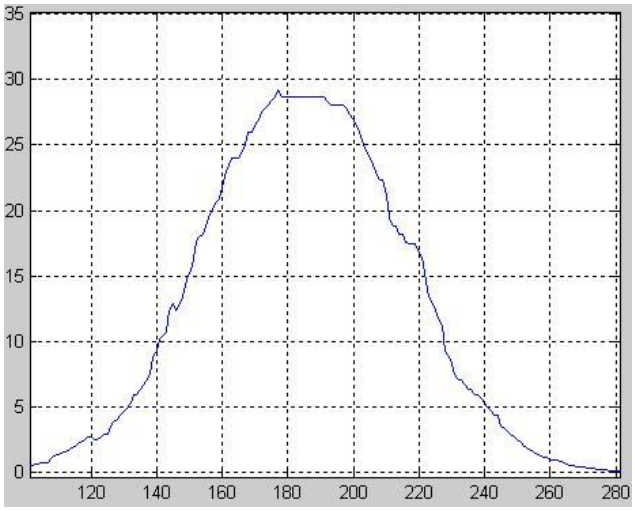

Рис. 3. Сигнал после отбраковки ложных выбросов алгоритмом медианы в скользящем окне (МСО) — искажена вершина

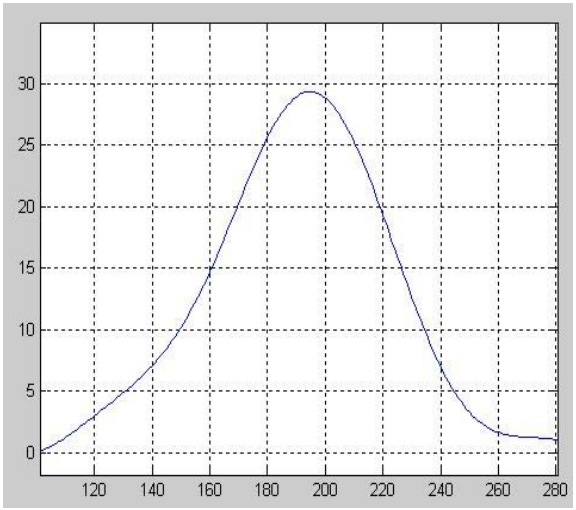

Рис. 4. Сигнал после отбраковки ложных выбросов цифровым фильтром Чебышева — искажена форма

Средние значения невязок от "идеального" сигнала для сигналов, обработанных разными алгоритмами, при различных количествах и амплитуде ложных выбросов

Параметры сигнала		Методы обработки			
% выбросов	Амплитуда выбросов	Мкмо	Мсо	Сглаживание	Фильтр Чебышева
25	20	2.34	2.97	9.22	4.88
25	30	2.52	2.98	9.47	12.31
25	40	2.30	2.93	10.19	12.33
35	20	2.32	2.95	6.51	10.24
35	30	2.42	2.50	9.00	12.96
35	40	2.31	3.20	12.70	12.27
40	20	2.34	2.68	6.21	10.31
40	30	2.40	2.70	8.09	11.43
40	40	2.74	2.84	13.32	13.25

АНАЛИЗ РЕЗУЛЬТАТОВ МОДЕЛИРОВАНИЯ

Результаты эксперимента (графики) — моделирование функции Гаусса, наложение шумов и выбросов, фильтрация зашумленной функции — представлены на рис. 1–4 и в таблице. Обозначены:

Мкмо — алгоритм минимизации квадрата медианного остатка,

Мсо — алгоритм медианы в скользящем окне,

Сглаживание — сглаживание в скользящем окне суммированием с "весами",

Фильтр Чебышева — цифровой фильтр Чебышева.

Из данных в таблице видно, что по сравнению с другими методами Мкмо имеет наименьшие значения средних значений невязок от "идеального" сигнала. Поэтому можно сделать вывод о том, что этот метод меньшего всего искажает "идеальный" сигнал.

Метод Мкмо оставляет на графике четко различимый пик, позволяющий дальнейший анализ причин его возникновения. Этот пик в дальнейшем может быть сглажен при повторном применении метода Мкмо.

ЗАКЛЮЧЕНИЕ

Рассмотренные выше алгоритмы, основанные на поиске минимума квадрата медианного остатка, являются универсальными для решения большинства задач обработки информации масс-спектрометрического изотопного анализа. Для ряда частных задач могут быть использованы более простые алгоритмы, описанные в работах [3, 4, 5, 6, 7].

СПИСОК ЛИТЕРАТУРЫ

1. *Rousseeuw P.J.* Least Median of Squares Regression // Journal of the American Statistical Association. 1984. V. 79. P. 871–880.
2. *Sirvidas S.I., Manoylov V.V.* The Software Outliers' Eliminator and Noise Smoother for Spectral Data // 7th International school-seminar on automation and computing in science, engineering and industry. РАН, МГУ. Ялта, 1996. P. 32.
3. *Манойлов В.В., Заруцкий И.В.* Отбраковка "выбросов" и оценка параметров масс-спектрометрических сигналов для прецизионного изотопного анализа // Научное приборостроение. 2002. Т. 12, № 3. С. 67–70.
4. *Манойлов В.В., Заруцкий И.В.* Алгоритмы первичной обработки масс-спектрометрических сигналов для прецизионного изотопного анализа // Вопросы атомной науки и техники. Серия "Техническая физика и автоматизация". Научно-технический сборник. Вып. 56. М.: Мин-во РФ по атомной энергии, Центральный научно-исследовательский институт информации и технико-экономических исследований в атомной науке и технике, 2002. С. 52–74.
5. *Манойлов В.В., Заруцкий И.В.* Алгоритмы обработки масс-спектрометрических сигналов для изотопного и химического анализа // Труды LVII Научной сессии, посвященной Дню радио. М.: Российское научно-техническое общество радиотехники, 2002. Т. 1. С. 274–277.
6. *Манойлов В.В., Аракелянц М.М., Чернышев И.В., Сердюк Г.И.* Программное обеспечение измерительно-вычислительной системы на основе IBM/PC-AT для определения возраста геологических образцов калий-аргоновым методом на масс-спектрометре МИ-1201ИГ // 12-й Международный симпозиум по проблемам модульных информационно-вычислительных систем и сетей. Тезисы докладов 3.7. РАН, МГУ. М., СПб., 1997. С. 43.
7. *Манойлов В.В., Аракелянц М.М., Чернышев И.В.* Программное обеспечение для определения изотопного состава аргона в автоматизированном комплексе на базе масс-спектрометра МИ1201ИГ // Научное приборостроение. 1999. Т. 9, № 4. С. 84–95.

Институт аналитического приборостроения РАН, Санкт-Петербург

Материал поступил в редакцию 19.05.2009.

THE INVESTIGATION OF ALGORITHMS OF OUTLIERS' ELIMINATOR IN MASS-SPECTRAL SIGNALS

V. V. Manoylov, I. V. Zarutsky

Institute for Analytical Instrumentation RAS, Saint-Petersburg

The results of investigation of some types of algorithms of outliers eliminator and smoothing in mass-spectral signal are discussed. The investigation was performed with the help of computational experiment on the

base models of mass-spectral signals, distorted by noise and outliers. The advantages of the modified Rousseeuw's method for mass-spectral analysis are shown. This method is based on the least median of squares regression.

Keywords: method for signal treatment, mass-spectometry, statistical data analysis, signal smoothing and filtration