

В. И. ВИНОГРАДОВ, В. Г. МУРАТОВ, В. И. ПЕТРОВА
(Ленинград)

ОРГАНИЗАЦИЯ АВТОМАТИЧЕСКОГО ОТБОРА ДАННЫХ В СПЕКТРОМЕТРИЧЕСКОЙ ИНФОРМАЦИОННОЙ СИСТЕМЕ

Характерными особенностями информационных систем для физических исследований являются рост количества информации и высокие требования по быстродействию и достоверности обработки. Поэтому при построении информационных систем подлежат анализу такие вопросы, как сокращение общего количества перерабатываемой в системе информации (количественный фактор, основанный на критерии ценности информации) [1, 2] и повышение надежности функционирования и общей достоверности обрабатываемых данных при минимуме вводимой в систему информационной и структурной избыточности (количественный фактор, не связанный с качественной стороной информации) [3].

Использование специальных схем отбора информации у источника не решает всех задач сокращения данных даже при однопараметрическом спектре. В случае трехпараметрического анализа с числом анализируемых каналов на каждый параметр 2^{10} общее количество каналов составляет 2^{30} . В подобных случаях из-за технических трудностей вместо интегрирующих накопителей используют неинтегрирующую память; при этом объем данных становится больше из-за необходимости непосредственной регистрации событий. Возникает задача разработки средств, позволяющих проводить многопараметрические исследования с отбором ценных областей в многомерных входных пространствах заданием минимального набора констант интересующих моделей ценной информации.

Процесс отбора ценной информации включает две задачи: 1) создание модели, определяющей область полезной информации либо задающей границу разделения входного пространства L возможных наблюдений на два подпространства — L_1 (источником может являться сообщение и фон) и L_0 (источником является только фон); 2) реализацию выбранного метода проверки гипотез при регистрации поступающих событий: H_0 (ценная информация отсутствует) и H_1 (ценная информация присутствует) для выработки решения R , задаваемого двумя значениями 0 и 1.

Реализация отбора соответственно состоит из двух этапов: предварительного накопления для определения модели и регистрации в соответствии с моделью «ценных» областей.

Известные методы отбора в процессе регистрации, например методы селекции [4], пригодные для случаев ярко выраженных пиков на

слабом фоне, приводят к удлинению эксперимента из-за большого отсева ценной информации. Метод парной селекции эффективен при обнаружении ценных областей для последующей регистрации по выбранным зонам или цифровым окнам. Принятые два последовательных события сравниваются по фиксированному числу старших разрядов, рассматриваемых как дескрипторные. В случае сравнения событие регистрируется как ценное. При неравенстве дескрипторов принимается следующее событие и осуществляется сравнение с дескриптором последнего из ранее принятых событий.

Отсев ценных событий приводит к снижению пропускной способности метода отбора, под которой понимается отношение числа зарегистрированных событий (k) к общему числу входных событий (M) для заданного распределения на входе системы.

Моделирование на ЭВМ «Минск-22» отбора k слов из M входных путем селекции из s по l машинных n -разрядных слов при их сравнении по r разрядам приведены в таблице. С ростом s увеличивается число регистрируемых событий.

M	l	s	$n-r$	k	$\frac{k}{M} \cdot 100 \%$
12000	2	2	0	43	0,36
	2	3	0	127	1,06
	2	2	2	133	1,24
	2	3	2	488	4,07
	2	6	0	608	5,07
	2	6	2	2408	20,07
40000	2	2	0	152	0,38
	2	3	0	446	1,11
	2	2	2	533	1,33

На рис. 1, а приведена модель анализируемого спектра. При отборе из 6 последовательно поступающих событий при их попарном сравнении без учета двух младших разрядов пропускная способность составляла около 0,2 (см. рис. 1, б). Однако с ростом s затрудняется техническая реализация таких систем и оказывается предпочтительнее алгоритмический отбор на ЭВМ. Последнее особо ощутимо в условиях высокоинтенсивного фона и слабой интенсивности сигнала.

Первая задача отбора аналогична задаче обнаружения сигнала, являющейся частной задачей различения сигналов, и ее решение сводится к двум гипотезам: H_0 и H_1 [5]. Реализация двух этапов возможна полностью как на ЭВМ, так и на специализированных информационных системах. Более эффективно реализация может быть осуществлена оптимальным комбинированием программно-аппаратных средств.

Формирование модели. Методика отбора многомерных событий аналогична методике отбора одномерных событий, поэтому в дальнейшем рассматривается одномерный спектр. В каждом интервале (x_{i-1}, x_i) гистограммы, называемом каналом, накапливается число поступивших событий v , среди которых могут быть как ценные события ($s \neq 0$), так и фон ($s=0$), причем

$$v(x) = v_{s \neq 0}(x) + v_{s=0}(x).$$

Возможны четыре варианта соотношений интенсивностей фона и полезного события:

$$v_{s \neq 0}(x) > v_{s=0}(x); \quad v_{s \neq 0}(x) < v_{s=0}(x);$$

$v_{s \neq 0}(x)$, $v_{s=0}(x)$ (относительно большие величины);

$v_{s \neq 0}(x)$, $v_{s=0}(x)$ (относительно малые величины).

Первый вариант соответствует распределению вектора, представленного узким пиком на относительно слабом фоне. Отбор ценной информации в этом случае решается проще всего по одному из разработанных методов, основанных на том, что ценность информации пропорциональна интенсивности составляющих спектра.

По предварительно отобранном k событиям формируется модель ценной информации спектра. При этом отношение статистически достаточного числа зарегистрированных событий к общему числу входных событий при заданном распределении на входе определяет пропускную способность системы.

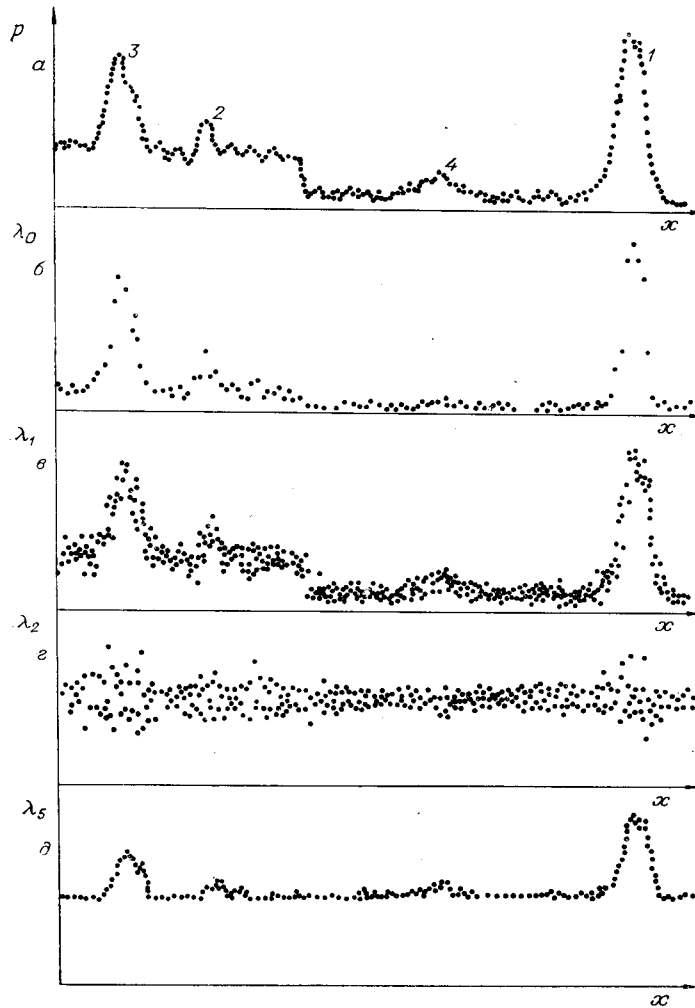


Рис. 1. Графики анализируемого спектра:

а — модель однопараметрического спектра $p(x)$; число каналов $k=x=512$;
 б — спектр отобранных событий $\lambda_0(x)$, селектуемых по $l=2$ из $s=6$ при
 $n-r=2$; $M=40\ 000$; $k=8095$; в — накопленный спектр $\lambda_1(x)=p(x)$ при $M=$
 $=12\ 000$; г — производная накопленного спектра $\lambda_2(x)=p'(x)$ при $M=12\ 000$;

$$д — модель спектра $\lambda_5(x) = \int_{x-T_0}^x p'(x) dx$; $T_0=35$; $\kappa=0$.$$

В случае ярко выраженного пика на низком уровне фона модель ценной информации достаточно просто задать в виде

$$\lambda_1(x) = p(x). \quad (1)$$

В реальных условиях часто встречаются спектры с большой интенсивностью фона. В этих условиях модель спектра эффективно задавать соотношением

$$\lambda_2(x) = p'(x), \quad (2)$$

что соответствует реализации отбора по производной рассматриваемого распределения [6], соответствующего информативности каналов [7]. Однако метод приближенного вычисления производной требует хорошей статистики (см. рис. 1). Практически оценка спектра часто осуществляется визуально экспериментатором на имеющихся средствах отображения данных в процессе его накопления.

С целью уменьшения чувствительности к высокоинтенсивным составляющим спектра и учета ослабления фона с увеличением x при задании модели можно воспользоваться отношением

$$\lambda_3(x) = p'(x)/p(x). \quad (3)$$

В условиях малой статистики для выделения известных распределений (пиков) в случайном спектре хорошие результаты дает также корреляционный метод обработки [8].

При необходимости регистрации отдельных подмножеств входного многомерного пространства с определенными функциональными связями параметров (например, заданные диапазоны отношений двух параметров) вычисляется модель функционального отбора

$$\lambda_4 = f(x, y, \dots). \quad (4)$$

Представляет интерес вычисление модели по производной $p'(x)$, полученной в условиях недостаточной статистики. Если взять интеграл

$$\lambda_5(x) = \int_{x-T_0}^x p'(x) dx \quad (\text{для } \lambda(x) > 0) \quad (5)$$

на участке длины T_0 и произвести обратное восстановление спектра во всех точках x на скользящем отрезке T_0 , то в полученной модели одновременно со сглаживанием на интервале $x \div (x - T_0)$ исчезает не только постоянная составляющая, но и плавные изменения функции распределения плотностей вероятностей (см. рис. 1). Граница области «ценной» информации может быть определена при этом из условия

$$\lambda(x) \geq \kappa,$$

где κ — выбранный порог.

Для использования ассоциативных режимов работы памяти модель задается набором дескрипторов по всему отбираемому подпространству событий. Для снижения числа операций, обработки входящих событий уменьшим число дескрипторов, переходя к контурному описанию модели, т. е. задавая минимальное число точек границы областей L_0 и L_1 .

Выбор границы можно осуществить заданием в качестве критерия оптимальности функции потерь; при этом каждой паре сигнала s и решения R ставится в соответствие величина, называемая потерей. Оптимизация системы обнаружения при этом производится по минимуму среднего риска $M[l(s, R)]$, $l(s, R)$ зависит только от наличия сигнала:

$$\begin{aligned} l(s=0, R=0) &= l_{00}; \quad l(s=0, R=1) = l_{01}; \\ l(s \neq 0, R=0) &= l_{10}; \quad l(s \neq 0, R=1) = l_{11}. \end{aligned}$$

Минимум среднего риска представляется в виде

$$M[l(s, R)] = q \{l_{00}p_{s=0}[R=0] + l_{01}p_{s=0}[R=1]\} + \\ + p \{l_{10}p_{s \neq 0}[R=0] + l_{11}p_{s \neq 0}[R=1]\} = \min, \quad (6)$$

где p — вероятность присутствия сигнала; q — вероятность отсутствия сигнала ($p+q=1$); $p_s(R)$ — вероятность принятия решения R при условии, что сигнал равен s .

Согласно критерию идеального наблюдателя, правильным решениям приписываем нулевые потери, а ошибкам обоих родов — потери, равные 1. Требование минимума среднего риска тогда сводится к общему виду

$$a_0p_{s=0}[R=1] + a_1p_{s \neq 0}[R=0] = \min. \quad (7)$$

Выбор коэффициентов a_0 и a_1 определяет порог $\kappa = a_0/a_1$, задающий границы подпространств L_0 и L_1 по выбранной модели $\lambda(x)$ из условия $\lambda(x) = \kappa$. Если основной задачей системы является сокращение общего объема регистрируемых данных при минимуме потерь ценных событий, то $a_0/a_1 \rightarrow 0$.

Цифровой отбор и сжатие данных. Второй этап отбора связан с реализацией проверки гипотез и выработкой решения о «ценности» события (R). Существующие методы ассоциативного накопления [9—11] требуют предварительного запоминания всего поля «ценных» дескрипторов. Более эффективно осуществляется отбор методом зонной селекции [12], однако эффективность метода зависит от распределения данных по зонам, которые для интегрирующей части памяти должны быть соизмеримы. При уменьшении размеров фигур и с ростом их числа зонный отбор приближается в пределе к ассоциативному. С увеличением размера элементарных фигур и снижением их числа уменьшается точность отбора «ценных» областей, но повышается пропускная способность системы, под которой понимается среднее число операций, необходимых для обработки каждого события, поступившего на вход. Поэтому было бы целесообразно обеспечить возможность отбора по произвольным многомерным конфигурациям в процессе приема данных, а вычисление модели ценной информации обеспечить программно по более совершенным алгоритмам с помощью ЭВМ.

С целью уменьшения числа точек, характеризующих границы «ценных» областей отбора, подпространство L_1 будем выделять областями (элементарными фигурами), границы которых параллельны осям, задающим все пространство. При этом каждую границу можно характеризовать отдельной точкой. Принимаемое многопараметрическое событие по всем параметрам должно проверяться на удовлетворение граничным условиям заданных ценных областей.

При регистрации ценных событий в интегрирующий накопитель элементарные фигуры должны быть эквивалентны по размерам. Нижние и верхние границы стыкуемых фигур обозначаются одной точкой, координаты которой определяются с точностью до значения младшего разряда истинного дескриптора (константы). Размеры фигур по параметрам выбираются произвольно. При регистрации данных в неинтегрирующий накопитель размеры отдельных элементарных фигур, покрывающих области ценной информации, выбираются произвольно. При этом константы находятся только на контуре интересующей области.

Совокупность решения по всем m параметрам определяет общее решение R_m о регистрации поступившего сообщения; при этом для $m=3$ возможна реализация (логических) соотношений подмножеств:

1) $VR^x R^y R^z = R_3$, если $k = l = s$ — число зон по параметрам x, y, z соответственно; (8)

2) $VR^x VR^y VR^z = R_3$, если $k \neq l \neq s$; (9)

3) $VR^x + VR^y + VR^z = R_3$, если $k \neq l \neq s$. (10)

Рассмотрим алгоритм работы и организацию системы цифровой фильтрации спектрометрических данных по произвольным конфигурациям, являющуюся подсистемой большой дискретной системы обработки данных [13].

Система цифровой фильтрации. Система цифровой фильтрации разработана с учетом недостатков и преимуществ существующих методов и средств сжатия. Цифровые данные, поступившие в групповой тракт приема в накопительные устройства от одной из выносных измерительных станций [14], могут поступать либо в интегрирующий, либо в неинтегрирующий накопитель как через систему фильтрации, так и непосредственно без отбора. Стандартные блоки оперативной памяти (ОП) автоматически могут замещать один из кубов ОЗУ ЭВМ «Минск-22». Средства отображения на экранах ЭЛТ, установленных как в центре обработки, так и у экспериментальных приборов, позволяют следить за ходом регистрации данных. Система состоит из управляющего устройства (УУ), блока сравнения (БС) по 3 параметрам x, y, z (10+10+12 разрядов), блока фильтрации (БФ), блока ускоренного поиска граничных констант (БУП), приемного регистра (ПР) и блока контроля (БК) (рис. 2).

Поступившее в цифровое приемное устройство (ЦПУ) сообщение передается автоматически на обработку в зависимости от номера абонента, определяемого 4 старшими разрядами, и выбранного режима.

Выделенные для хранения констант 1024 ячейки памяти предоставляют возможность каждому из 12 абонентов заносить до 64 констант по каждому из 3 параметров. Граничные константы могут быть введены программно вручную с пульта ЭВМ либо с помощью телетайпа, расположенного рядом с выносным оборудованием.

Разделение памяти между абонентами осуществляется в соответствии с разбиением адресов констант: $MA BC$, где M — номер группы абонентов (HA_1, HA_2); A — номер абонента в восьмеричной системе (1, 2, ..., 12), B, C — любые восьмеричные числа.

В системе предусмотрены 2 входных групповых тракта: а) первый ($M=7$) принимает данные в первый приемник и по номеру A направляет информацию в анализатор буферной HA_1 (16 000 каналов) либо в буферный накопитель, емкость которого 3072 слова по 37 разрядов; б) второй ($M=6$) принимает данные во второй приемник от второй группы абонентов, а по номеру A направляет их автоматически в анализатор буферной HA_2 (8000 каналов) либо в тот же накопитель буферной (НБ).

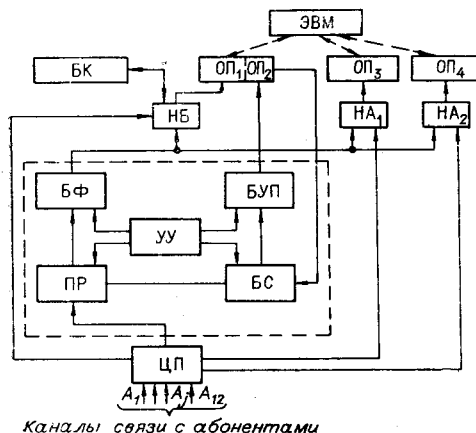


Рис. 2. Блок-схема цифровой фильтрации многомерных случайных событий.

Если требуется сжатие данных, то система фильтрации в разделение времени включается в любой из перечисленных трактов обработки данных. Кроме того, ячейки 6000—5077, 6700—6777 и 7700—7777 структурно выделены для служебных целей и могут использоваться произвольно, а ячейки 7000—7077 предоставлены для организации схемного динамического контроля работоспособности системы, периодически проверяющего на правах отдельного (нулевого) абонента весь тракт фильтрации, включая ускоренный поиск, сравнение чисел, а также тракты хранения и чтения чисел в накопителе. Сбой констант обнаруживается схемно и может восстанавливаться с помощью программы ЭВМ в автоматическом режиме. При необходимости проведения двухпараметрического или однопараметрического анализа с большим числом каналов используется режим отбора со сжатием для дальнейшей регистрации в одном из накопителей анализаторных (НА). При трехпараметрических исследованиях используется режим отбора с последующей регистрацией в буферный накопитель, причем возможно автоматическое занесение вместо одного из параметров признака номера зоны для упрощения дальнейшей машинной сортировки данных.

Сообщение при фильтрации с регистрацией в НА по каждому параметру разбивается на 2 части: 1) дескриптор, содержащий до $k \times 2$ разрядов, 2) номера каналов, меняющиеся по желанию до $k \times 2$ каналов, где $k=2, 3, 4, 5$.

Структура сообщения после фильтрации может быть по желанию экспериментатора установлена на коммутационной панели, на которую без изменений поступают 5 младших разрядов параметров x и y , а старшие разряды заменяются условным номером дескриптора, получаемого с БУП. На абонентских экранах изображаются при этом по желанию любые комбинации многомерных сечений по параметрам и дескрипторам накапливаемого в НА спектра.

В системе достаточно реализации решения по первым двум соотношениям (9) и (10), так как они позволяют задать любые конфигурации в исследуемом пространстве. При фильтрации по первому соотношению поиск констант производится по верхним границам, а проверка удовлетворения события требуемым условиям — по нижним. Константы при этом могут быть не упорядочены. В четных адресах выделенной зоны запоминающего устройства всегда хранятся нижние границы областей ценной информации, а в нечетных — верхние. При линейном поиске перебираются последовательно нечетные адреса. Поиск осуществляется по одному из параметров, а проверка условий во всех параметрах производится одновременно.

При фильтрации по второму соотношению поиск констант осуществляется по нижним границам, а проверка гипотезы для выработки решения — по верхним. БУП работает при упорядоченном расположении констант методом двоичного взвешивания. Таким образом, адрес нижней границы будет найден за $\lg_2 k$ шагов работы БУП, после чего проверяется одно условие на удовлетворение верхней границе, константа которой хранится в памяти по адресу, на единицу большему адресу последней нижней границы.

Система в настоящее время успешно эксплуатируется и показала широкие возможности по решению различных задач в спектрметрических исследованиях. Использование общих линий связи от станций и различных типов каналов прямого доступа к блокам памяти, которые автоматически замещают МОЗУ ЭВМ «Минск-22», а также гибкая система отбора и сжатия данных, работающая на любой из этих трактов обработки, значительно расширили функциональные возможности при

высоком уровне автоматизации организации системы и позволили проводить широкий класс спектрометрических исследований.

ЛИТЕРАТУРА

1. Сокращение объема измерительной информации. Труды I симпозиума. Новосибирск, «Наука», 1968.
2. В. И. Рабинович, М. П. Цапенко. Информационные характеристики средств измерения и контроля. М., «Энергия», 1968.
3. В. И. Виноградов. Вопросы оптимизации дискретных информационных систем. Препринт ФТИ-275. Л., 1970.
4. Z. Hofmann. Statistical analysis of the method of pairselection in content addressable system. Doctoral — thesis, Swierk, 1966.
5. П. П. Сысоев. Оценки параметров, обнаружение и различение сигналов. М., «Наука», 1969.
6. E. Gatti. In Automatic acquisition and reduction of nuclear data.— Proc. Conf. EANDC, Karlsruhe, Gesellschaft für Kernforschung, mbH, 1964.
7. А. П. Нехай. Метод ассоциативного анализа, основанный на определении производной от измеряемого распределения.— Труды VI конференции по радиоэлектронике. Алушта, 1968.
8. W. W. Black. Application of correlation techniques to isolate structure in experimental data.— NIAM, 1969, v. 71.
9. G. C. Best, S. A. Hickman. A Contents Addressable Analyser for Million Channels.— IEEE Trans. on Nucl. Science, 1966, NS-13, № 13.
10. Ch. Goodman. Automatic acquisition and reduction of nuclear data.— Proc. Conf. EANDC, Karlsruhe, Gesellschaft für Kernforschung, mbH, 1964.
11. Р. Г. Офенгенден и др. Дискретная система для измерений многомерных распределений вероятностей. Киев, «Наукова думка», 1969.
12. I. N. Hooton. Zone Selection in Multiparameter Analysis.— IEEE Trans. on Nucl. Science, 1966, v. NS-13, № 13.
13. В. И. Виноградов, В. И. Кадашевич, В. И. Петрова. Многоканальный цифровой ввод данных в ЭВМ при помощи специализированных накопительных систем.— Приборы и техника эксперимента, 1970, № 4.
14. В. И. Кадашевич, В. И. Виноградов, В. В. Марченков, А. П. Нехай. Выносная измерительная станция центра спектрометрических измерений. Препринт ФТИ-199. Л., 1969.

*Поступила в редакцию
14 июня 1971 г.*