ISSN 0868–5886 НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2000, том 10, № 1, с. 35–43

ОРИГИНАЛЬНЫЕ СТАТЬИ

УДК 621.391.14

© М. М. Нестеров, В. Н. Трифанов, В. Н. Данилов

НЕСТАНДАРТНЫЙ АНАЛИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ САМООРГАНИЗУЮЩИХСЯ ТЕХНОЛОГИЙ

Излагаются основы и общие принципы нестандартного анализа данных с использованием самоорганизующихся технологий анализа и проявления скрытой организованности и периодичности данных в иерархическом (ультраметрическом) режиме. Частично затронуты технологии лексикографического анализа данных, их конструктивного анализа и статистической фильтрации с целью применения в современном приборостроении.

ВВЕДЕНИЕ

В период развития компьютерных технологий анализа данных в режиме реального времени актуальным становится не только качественный рост быстродействия и памяти компьютеров, но и быстрые, устойчивые и эффективные процедуры обработки, анализа и интерпретации данных. Традиционные стандартные технологии статического, динамического и статистического анализов данных в большинстве своем являются ресурсоемкими. Как правило, они эффективно работают на массивах малой мощности, описывающих сравнительно гладкие предсказуемые процессы с низким уровнем шума. Однако на практике требуется обрабатывать большие и сверхбольшие массивы данных со слабой предсказуемостью и высоким уровнем шума.

Для решения этих запросов практики в режиме реального времени в последнее время разрабатываются эффективные технологии анализа данных на базе нестандартных конструктивных, лексикографических и статистических процедур обработки, фильтрации, анализа и интерпретации данных. Достигается это за счет рекурсивных процедур самоорганизации и ультраметрической организации данных.

КРАТКИЕ ОСНОВЫ И ОБЩИЕ ПРИНЦИПЫ НЕСТАНДАРТНОГО АНАЛИЗА ДАННЫХ

Выбор конкретной технологии конструктивного анализа реальных сигналов определяется конкретным классом измеряемых процессов, способом их измерения, степенью разработки и реализуемости технологий обработки, отображения и интерпретации сигналов в соответствии с поставленными целями и требованиями к разрешающей способности обнаружения, идентификации и распознавания выделяемых информационных объектов, а также ко времени решения поставленных задач.

Два характеристических условия конструктивного анализа: хорошая физическая интерпретируемость технологических анализа и эффективность их выполнения определяют открытый класс технологических решений, который постоянно пополняется новыми исследованиями.

Традиционные методы спектрального, дисперсионного и корреляционного анализов сигналов дают удовлетворительные результаты при обработке стационарных сигналов в условиях малого шума. Однако при обработке нестационарных сигналов с высоким уровнем шума они практически перестают работать [1]. Различные технологии время-частотной и динамической фильтраций часто наводят ложные сигналы (фантомы) и тем самым затрудняют обработку реального сигнала. Но главный недостаток всех этих методов — их большая трудоемкость.

Некоторые преимущества в эффективности обработки сигналов исследователи получили, перейдя от спектральных методов к локально связным методам типа методов конечных и ограниченных элементов и опытов [2, 3]. Однако и здесь в задачах большой размерности возникают проблемы трудоемкости и устойчивости, которые до сих пор не преодолены.

Хорошие результаты по повышению устойчивости решений дают методы интегральных соотношений, разрабатываемые школой Дородницина и Белоцерковского, однако и здесь проблема трудоемкости обработки остается неразрешимой [4].

Интуитивно становится ясным, что быстрые процедуры обработки сигналов могут быть реализованы только на выборках, если эти выборки не очень большие и информативно представительны. В этом смысле выборки, построенные на множестве экстремальных точек сигнала, являются практически минимальными, так как множество экстремумов является множеством меры нуль в полном множестве точек сигнала.

Технология нестандартного анализа, предлагаемая в данной работе, основана на выборках та-

кого рода. Эти технологии имеют примерно линейный рост трудоемкости от объема сигнала, тогда как традиционные методы, основанные на анализе разных квадратичных форм и парных сравнениях, имеют кубический и более высокий рост трудоемкости от размерности решаемой задачи.

Однако в условиях шума когерентная совокупность экстремальных точек становится неустойчивой. В этом случае требуется разработка дополнительных технологий фильтрации сигнала. Простейшие способы фильтрации — это анализ частичных сумм [5]. Технологии организации частичных сумм разнообразны. Они могут быть локально связны и несвязны, статистические, скользящие и динамические, взвешенные и невзвешенные, декларативные по протекции, самоорганизующиеся и смешанные, робастные и локусные и т.д. Этот классификационный перечень открыт, и он будет пополняться новыми и новыми технологиями.

В некотором смысле они являются прототипами метода интегральных соотношений Дородницина—Белоцерковского [4]. Этот метод обладает хорошей физической интерпретируемостью и эффективностью анализа [6]. В данной работе этот метод является одним из основных.

Главным для всех этих технологий является инвариантность представления сигнала в исходном пространстве и пространстве замещающих точек. Известно, что со всякой совокупностью инвариантов связаны симметрии и законы сохранения. Таким образом, метод замещающих точек связан с анализом проблемы симметрии—асимметрии и ее проявления в исследуемом сигнале.

Так как в методе замещающих точек исходный процесс представлен в компактифицированной форме с точностью до инвариантов, то здесь вскрываются большие резервы по компрессии и декомпрессии исследуемого сигнала. Как правило, трудоемкость обработки компактифицированного сигнала растет примерно линейно с ростом его объема.

В условиях зашумленного сигнала возникает проблема разделения его детерминированной и индетерминированной составляющих. Это одна из центральных проблем в технологиях фильтрации сигнала. Граница между этими составляющими неопределенна и проблематична. Здесь требуется некоторый прагматичный подход. Один из таких подходов разрабатывается нами в технологиях статистической фильтрации. Это одна из разновидностей технологий метода замещающих точек, в котором инвариантами выступают статистические моменты [5]. Реализуемые в режиме самоорганизации, они удовлетворяют критериям полноты, эффективности и устойчивости описания сигнала в пространстве замещающих точек. При этом технология описания такова, что детерминированная и индетерминированная составляющие сигнала ортогональны. Следовательно, статистический фильтр, построенный таким образом, является помехоустойчивым [5, 7, 8].

Следующей отличительной особенностью совокупности выборочных точек является возможность их иерархической организации. В этой технологии множества, построенные по инвариантам первичного замещения, вновь замещаются по инвариантам вторичного, третичного и так далее замещения. Такую организацию физики называют ультраметрической. Именно она, эта технология, позволяет практически реализовать условия Колмогорова линейного роста трудоемкости обработки сложного сигнала с ростом его размерности.

Основное внимание в следующих разделах будет уделено высокоэффективным методам когерентного анализа данных. Когерентное суммирование обладает рядом уникальных свойств. Во-первых, при когерентном суммировании отношение сигнал/шум растет пропорционально числу когерентно суммируемых пакетов (кластеров). Из этого следует, что даже при слабом сигнале, когда отношение сигнал/шум для одного пакета очень мало и ниже порога разрешающей способности обнаружения сигнала имеющимися измерительными средствами, то при соответствующем числе суммируемых пакетов этот порог будет преодолен, сигнал обнаружен. Во-вторых, когерентное суммирование в связи с этим свойством является помехоустойчивым и селективным с точки зрения различения обнаруженных сигналов. В-третьих, сложность когерентного суммирования растет линейно, пропорционально числу суммируемых пакетов. Именно этим свойством достигается линейная простота алгоритма селекции в соответствии с теоремой Колмогорова.

КОНСТРУКТИВНЫЙ АНАЛИЗ ПЕРИОДИЧЕСКИХ ПРОЦЕССОВ

Конструктивный анализ, по определению, предполагает выполнение двух основных требований [9]:

1. Нестандартное построение фрагментов, элементов и процедур анализа, основанное на их физической интерпретируемости, инструментальной измеримости и технологической локализуемости.
2. Разработка нестандартных процедур эффективного обнаружения, проявления, распознавания и классификации и анализа как проявленных, так и непроявленных объектов анализа, описываемых выборочными кластерами данных с измеримыми признаками их локализации.

Конструктивный анализ данных базируется на самих данных, организованных в последовательность или процесс по характерному для них пара-

метру порядка. При этом даже не требуется вводить базисные функции описания процесса, так как эти функции конструируются самими данными.

В технологиях нестандартного анализа данных и его разновидностях, в том числе и конструктивного анализа, на первый план выдвигаются проблемы объективной группировки данных в организованные локальные кластеры, их повторяемость и периодичность с выделением измеряемых признаков их обнаружения, локализации, организации и распознавания [10, 11]. Рассмотрим некоторые особенности конструктивного анализа случайных процессов при обнаружении их скрытых периодичностей и их взаимной организованности.

Пусть данные упорядочены параметром порядка t в некоторый статистический процесс $x(t)$. В общем виде случайный процесс при наличии в нем скрытой периодичности можно представить выражением

$$x_1(t) = \chi_1(t) f_1(t + n_1 T_1(t)), \quad n_1 = [1, N_1] \subset Z, \quad (1)$$

где $\chi(t)$, $T(t)$ — случайные величины, зависящие от параметра t, f — аппроксимирующая функция.

Под скрытой периодичностью понимается функция

$$\overline{x_1}(t) = \chi_1 f_1(t + n_1 T_1), \quad (2)$$

где $\chi_1 = M(\chi_1(t))$, $T_1 = M(T(t))$, M — операция математического ожидания по параметру t.

В общем виде исходный процесс первого порядка $x_1(t)$ и его аппроксимирующий процесс скрытой периодичности первого порядка $\overline{x_1}(t)$ отличаются друг от друга. Невязку такой аппроксимации можно рассматривать как случайный процесс второго порядка

$$x_2(t) = x_1(t) - \overline{x_1}(t) = \chi_2(t) f_2(t + n_2 T_2(t)), \quad n_2 = [1, N_2] \subset Z. \quad (3)$$

Этот процесс содержит в себе процесс скрытой периодичности второго порядка

$$\overline{x_2}(t) = \chi_2 f_2(t + n_2 T_2), \quad (4)$$

где $\chi_2 = M(\chi_2(t))$, $T_2 = M(T_2(t))$.

Такую процедуру выявления скрытых периодичностей можно продолжить до требуемой разрешающей способности описания исходного процесса как по точности, так и по прогностической устойчивости описания. На шаге k такого описания имеем процесс порядка k

$$\begin{aligned} x_k(t) &= x_{k-1}(t) - \overline{x_{k-1}}(t), \\ x_k(t) &= \chi_k(t) f_k(t + n_k T_k(t)), \end{aligned} \quad (5)$$

$$\overline{x_k}(t) = \chi_k f_k(t + n_k T_k), \quad (6)$$

где $\chi_k = M(\chi_k(t))$, $T_k = M(T_k(t))$.

Такой способ описания является спектральным, так как он аппроксимирует на всем множестве параметры порядка $t \in J_t$. Спектральный способ описания изначально присущ стандартному классическому анализу, который ограничивается описанием первого порядка одной аппроксимирующей функцией или разложением в аппроксимирующие ряды базисных функций.

В конструктивном анализе скрытых периодичностей наблюдается иерархическая (ультраметрическая) рекурсия. Это одна из принципиальных особенностей нестандартного спектрального анализа. Качество конструктивного описания процесса на каждом шаге ультраметрической рекурсии зависит от выбора амплитуд $\chi_k(t)$ и весовых периодических функций $f_k(f + n_k T_k(t))$, а также от способа определения (χ_k, T_k) на множестве параметра порядка J_t.

В классическом анализе выбираются, как правило, базисные функции либо аппроксимирующие функции типа полиномов Лагранжа. В конструктивном анализе эти функции вообще могут быть не обозначены. Они могут быть «выращены» в процессе самоорганизующего обучения либо частично, либо полностью.

Полное описание возникает тогда, когда реальный процесс является чисто однопериодическим на каждом рекурсивном шаге. Это чрезвычайно редкое событие. Поэтому, как правило, в конструктивном анализе встречается частичное описание этих функций по некоторой выборке на периоде T, по которой строятся (конструируются) распознающие признаки каждого периода nT. Именно по сходству (повторяемости) этих признаков все множество параметра порядка J_t разбивается на совокупность периодов nT, $n \in [1, N] \subset Z$. Характер представительных частичных выборок и конкретный вид отличительных признаков не предопределен и является эвристическим. Их множество определяет конкретную технологию конструктивного анализа.

В другом локальном способе конструктивного анализа все множество параметра порядка J_t разбивается на последовательную совокупность непересекающихся множеств J_{nt}, причем

$$J_1 + J_2 + \ldots + J_N = J_t, \quad (7)$$

где $J_n = [t_n, t_{n+1}], \forall n \in \overline{1, N}$. В пределах каждого локального интервала амплитуда и период сигнала считаются постоянными. При этих условиях полное описание процесса имеет вид:

$$\chi(t) = \sum_{1}^{N} \chi_n f_n(t_n + \tau_n), \qquad (8)$$

где t_n — начало интервала J_n, $\tau_n = t - t_n$, $t \in J_n$, t — внутренний параметр порядка интервала J_n, $t_{n+1} - t_n = T_n$ — период интервала J_n.

Такое описание является локальным. Оно интенсивно разрабатывается в методах конечных элементов. Обычно в методах конечных элементов множество параметра порядка J_t равномерно разбивается на непересекающиеся множества J_n, $n = \overline{1, N}$, за редким исключением. Это вызвано тем, что в ставших уже классическими методах конечных элементов до сих пор нет конструктивных методов их дифференциации.

В конструктивном анализе эта проблема решается естественным образом путем конструирования измеримых признаков, выборочных точек каждого периода. Тогда интервалы повторяемости этих признаков разбивают естественным образом все множество J_t на его подмножества J_n со своими амплитудами χ_n, периодами T_n и функциями формы $f_n(t_n + \tau_n)$.

Выше уже отмечалось, что отличительные признаки скрытых периодов в конструктивном анализе носят эвристический характер. Основное требование к ним сводится к их физической интерпретируемости и измеряемости. Наиболее естественная и простая технология конструктивного анализа сводится к обнаружению границ локальных интервалов J_n по некоторым признакам [12]

$$\begin{cases} \chi(t_n) - \chi(t_{n+1}) = 0, \\ \chi'(t_n) - \chi'(t_{n+1}) = 0; \end{cases} \qquad (9)$$

$$\begin{cases} \chi(t_n) - \chi(t_{n+1}) = d\chi, \\ \chi'(t_n) - \chi'(t_{n+1}) = 0; \end{cases} \qquad (10)$$

$$\begin{cases} \chi(t_n) - \chi(t_{n+1}) = 0, \\ \chi'(t_n) - \chi'(t_{n+1}) = d\chi'; \end{cases}$$

$$\begin{cases} \chi(t_n) - \chi(t_{n+1}) = d\chi, \\ \chi'(t_n) - \chi'(t_{n+1}) = d\chi'. \end{cases} \qquad (11)$$

Условие (9) строго фиксирует уровень и производную уровня в начале каждого интервала $t_n \in J_n$, а по их повторяемости определяются последующие интервалы t_{n+1}, t_{n+2}, … , t_N. Эта технология конструктивного анализа пригодна для поиска скрытых периодичностей в сингулярных сигналах со слабым регулярным шумом.

При наличии регулярного шума достаточно высокого уровня ожидать точного сходства признаков начала интервалов разбиения множества J_t на его подмножества J_n, $n = \overline{1, N}$, не приходится. Признаки в этих условиях могут совпадать только с некоторой погрешностью $d\chi$, $d\chi'$, размер которой задается из прагматических целей, балансирующих точность и устойчивость их измерения. Технологии такого типа характеризуются признаками (10–11).

Скрытая периодичность на каждом шаге конструктивного анализа проявляется в результате осреднения амплитуд χ_n и периодов T_n по всему множеству интервалов $J_n \in J_t$ или по его частичной выборке.

Первый случай пригоден при анализе сингулярных сигналов с постоянным периодом ($T_n = T$) и с малым уровнем шума. В этом случае сигнал и форма скрытой периодичности будет иметь вид

$$\begin{cases} \chi(t) = \chi f(t_n + \tau_n), \\ \tau_n = t - t_n, t \in J_n, \\ \chi = M(\chi), T = M(T_n), \\ f = M(f_n), n = 1, N. \end{cases} \qquad (12)$$

При наличии регулярного шума достаточно высокого уровня периоды T_n интервалов J_n могут отличаться значительно друг от друга. Тогда при наложении сигналов этих периодов друг на друга при совмещении их начал концы этих периодов не будут совпадать. Возникнет неопределенность осреднения по параметру порядка t.

Среди множества технологий разрешения этой неопределенности выделим две простейшие. В одной из них предлагается перейти от абсолютного интервального времени $\tau_n = t - t_n$ к относительному

$$\overline{\tau_n} = \frac{t - t_n}{t_{n+1} - t_n} = \frac{t - t_n}{T_n}, \quad t \in J_n. \qquad (13)$$

В этом случае все локальные относительные интервалы времени принимают значение на отрезке

$$\overline{\tau_n} \in [0,1].$$

Их начала и концы совпадают. Поэтому становится возможным осреднение по всему множеству интервалов $J_n \in J_t$.

Во втором случае весь диапазон изменения интервалов

$$T_n \in [\min T_n = T_m, \max T_n = T_s], \quad T_n \in [T_m, T_s],$$

разбивается на ограниченное число классов $T_n \in I_k$, если

$$T_n \in \left(T_m + \frac{T_s - T_m}{K} \cdot k, T_m + \frac{T_s - T_m}{K} \cdot (k+1) \right). \qquad (14)$$

Считается, что все периоды $T_n \in I_k$, принадлежат одному классу I_k и характеризуются одной скрытой периодичностью. Осреднение для ее проявления производится по всему множеству периодов этого класса в относительном времени

$$\overline{\tau_n} = \frac{t - t_n}{\tau_n},\ t \in J_n,\ \tau_n \in I_k . \qquad (15)$$

Начала и концы периодов осреднения в относительном времени совпадают. Поэтому осреднение будет выполнено корректно. Однако, вместо одного среднего периода возникает целый спектр периодов I_k, которым соответствует спектр скрытых периодичностей.

В стандартном регрессионном анализе для поиска скрытых периодичностей строят регрессионные суммы второго порядка

$$R(t_n) = \sum x(t) y(t - t_n), \quad t \in J_t, \quad n = \overline{1, N} . \qquad (16)$$

В этих суммах исходный сигнал умножается на себя со сдвигом на t_n.

В рассмотренных выше технологиях конструктивного анализа составляются суммы первого порядка. Преимущества такого подхода очевидны.

Рассмотрим еще одну технологию нестандартного конструктивного анализа скрытых периодичностей, в котором признаками сходства интервалов разбиения являются признаки экстремумов, то есть признаки максимумов и минимумов. В каждой точке разбиения множества J_t возьмем еще два соседних отсчета сигнала слева и справа. В результате получим тройку:

$$t_{n1} \le t_n \le t_{n2} . \qquad (17)$$

В точке максимума соблюдаются признаки

$$t_n \ge t_{n1}; t_n \ge t_{n2} . \qquad (18)$$

В точке минимума имеем другие признаки

$$t_n \le t_{n1}; t_n \le t_{n2} . \qquad (19)$$

Разбиение по максимумам дает одно разбиение множества J_t, а по минимумам другое. В первом случае период T_n расположен между максимумами t_n, t_{n+1}, во втором случае — между минимумами t_n, t_{n+1}. Считается, что такие разбиения весьма эффективны, так как множество точек экстремума есть множество меры нуль по сравнению со всем множеством J_t, то есть это достаточно крупные разбиения и в то же время весьма информативные.

В этих технологиях исходный процесс 1

$$x_1(t) = \chi_1(t) f_1(t + n_1 T_1(t)), \quad n_1 \in [1, N] \subset Z$$

аппроксимируется некоторой функцией, например функцией Лагранжа, проходящей через экстремальные точки (максимумов или минимумов)

$$\overline{x_1}(t) = \bar{I} \chi_n f_n(t) . \qquad (20)$$

Аппроксимация выполняется либо в спектральном, либо в локальном исполнении. Одна аппроксимирующая функция является огибающей максимумов первого порядка, а другая — огибающей минимумов первого порядка. Такие огибающие, с одной стороны, выделяют интервалы периодичности (максимумов или минимумов), которые анализируются по вышеизложенным технологиям, а, с другой стороны, сами становятся предметами анализа на следующем рекурсивном шаге*.

Таким образом строится вложенная иерархическая (ультраметрическая) совокупность огибающих, каждая из которых является разбиением на скрытые периодичности, технология обработки которых изложена выше.

ЛЕКСИКОГРАФИЧЕСКИЙ АНАЛИЗ ДАННЫХ

Лексикографический анализ также относится к нестандартным процедурам обработки данных и является, в некотором смысле, дальнейшим развитием конструктивного анализа.

Впервые систематическую разработку этой технологии на примере анализа кардиограммы сердца выполнил Ю.И. Сенкевич в своей диссертации («Разработка математической модели и алгоритмов определения функционального состояния биологических объектов», 1998 г.). Особенность такой технологии была продиктована специфической структурой кардиоцикла [13, 14].

$$\begin{cases} P, Q, R, S, T, P; \\ t_n, t_{n+1}, t_{n+2}, t_{n+3}, t_{n+4}, t_{n+5}; \\ T_{n1}, T_{n2}, T_{n3}, T_{n4}, T_{n5}, \end{cases} \qquad (21)$$

где P — максимум, Q — минимум, R — второй максимум, S — второй минимум, T — третий максимум. Обычно R-максимум наиболее ярко выражен, поэтому в медицинской практике кардиоцикл измеряют R-R ритмами. В этой структуре периодический сигнал имеет сложную конструкцию с амплитудами в экстремальных точках

* Даже в простейшем случае выделения синусоиды из шума, трудоемкость стандартного спектрального анализа растет быстрее, чем в кубической степени, поскольку и в этом случае требуется проводить Фурье-анализ всех спектральных компонент исследуемых сигналов для того, чтобы выделить искомую компоненту. В то же время трудоемкость построения огибающей (соответствующей искомой синусоиде) по экстремальным точкам, например методом триад, который изложен в конце следующего раздела, растет линейным образом, поскольку экстремальные точки составляют множество меры нуль во множестве точек исследуемого сигнала.

$$x(t_n), x(t_{n+1}), x(t_{n+2}), x(t_{n+3}), x(t_{n+4})$$

и с интервалами между ними

$$T_{n1} = t_{n+1} - t_n, T_{n2} = t_{n+2} - t_{n+1}, T_{n3} = t_{n+3} - t_{n+2},$$
$$T_{n4} = t_{n+4} - t_{n+3}, T_{n5} = t_{n+5} - t_{n+4}.$$

Эти десять характеристик (пять амплитуд и пять интервалов) являются признаковыми. В лексикографическом анализе кардиоциклов эти десять характеристик являются знаком (буквой), а вся кардиограмма — последовательностью букв.

Более содержательно: знак (букву) на цикле $T_n = t_{n+5} - t_n$ можно представить матрицей

$$\mathbf{A}(t_n) = \left|\begin{array}{c|c|c|c|c} x_{11} & T_{12} & T_{13} & T_{14} & T_{15} \\ x_{21} & x_{22} & T_{23} & T_{24} & T_{25} \\ x_{31} & x_{32} & x_{33} & T_{34} & T_{35} \\ x_{41} & x_{42} & x_{43} & x_{44} & T_{45} \\ x_{51} & x_{52} & x_{53} & x_{54} & x_{55} \end{array}\right|, \qquad (22)$$

где

$$x_{11} = P, x_{22} = Q, x_{33} = R, x_{44} = S, x_{55} = T,$$
$$x_{km} = x_{kk} - x_{mm}, T_{km} = T_m - T_k.$$

В аналоговом представлении таких букв несчетное множество, но их можно сократить, если вместо сигнала ставить знак сигнала

$$x_{km} = sign(x_{km}); \quad T_{km} = sign(T_{km}). \qquad (23)$$

Количество букв такого алфавита значительно меньше. Их количество зависит от типа процесса. Для неорганизованных хаотических процессов алфавит букв бесконечен. По мере роста организованности алфавит уменьшается до некоторого уровня насыщения, характеризуемого соответствующим уровнем организованности процесса, определяемого пропорциями между его сингулярной (детерминированной, предсказуемой) и регулярной (хаотической, шумовой, непредсказуемой) составляющими [15, 16].

Таким образом, полностью или частично организованные процессы имеют конечные алфавиты [17]. Каждая буква алфавита удовлетворяет двум требованиям: 1) она отличается от других букв (признак разнообразия); 2) она повторяется в языке данного процесса (признак стабильности).

Последовательности букв $\mathbf{A_1}$, $\mathbf{A_2}$, … , $\mathbf{A_k}$ составляют слова языка. Слова языка, так же как и буквы, удовлетворяют требованиям разнообразия и стабильности. Количество слов языка также зависит от организованности процесса. Для хаотического процесса их бесконечное множество. Для организованного процесса множество слов ограничено. Чем более стабилен процесс, тем меньше и слов, и букв.

Возможны вообще крайние ситуации: для предельно стабильного процесса (абсолютно ритмичного самоподобного) существует только одна буква и только одно слово. Коль скоро реальные процессы полистабильны, то количество букв и слов таких процессов более разнообразно. Хаос нарушает стабильность и увеличивает разнообразие букв и слов. Полистабильность увеличивает разнообразие букв и слов при сохранении уровня организованности процесса. Эти две, казалось бы, похожие по форме тенденции, но полярные по содержанию отличаются друг от друга характером последовательности слов, которые составляют фразы языка.

Для хаотических неорганизованных процессов количество фраз неограничено. Организованные процессы имеют ограниченный стабильный набор фраз. Каждая фраза характеризует режим процесса, соответствующий определенному стабильному состоянию. Переход от одной фразы к другой связан с переходом процесса из одного стабильного состояния в другое, от одного режима функционирования к другому [5, 15, 16]. По набору букв, слов и фраз, а также по их последовательности можно судить о микро-, макросостояниях процесса и его режимах. Все это используется для диагностики состояния и режима процесса в его статике, динамике и для статистики. Медленные, эволюционные изменения алфавита, словаря и фраз дают представление о возрастном изменении в состояниях и режимах процесса в его эволюционном развитии.

Теперь рассмотрим технологию сравнений букв на предмет выявления их различия и схожести. Изначально каждая буква характеризуется вектором амплитуд $\boldsymbol{\chi} = (\chi_1, \chi_2, \chi_3, \chi_4, \chi_5)$ и вектором интервалов между экстремальными амплитудами $\mathbf{T} = (T_1, T_2, T_3, T_4, T_5)$. Рассмотрим два слова с амплитудами x, y и с интервалами T^x и T^y и из их компонент составим матрицу

$$\mathbf{A} = \left|\begin{array}{c|c} \chi_{mk} & 0 \\ \hline 0 & T_{mk} \end{array}\right|,$$

где $\chi_{kk} = \chi_k$, $T_{kk} = T_k^x$, $\chi_{mk} = \chi_m - \chi_k$,

$$T_{mk} = T_m^x - T_k^x, \quad m > k, \ \chi_{mk} = y_k - y_m,$$
$$T_{mk} = T_k^y - T_m^y, \ k > m, \text{ а } \mathbf{A} \text{ — матрица } 10\times10.$$

Если буквы x, y одинаковы, то, по построению, матрица $\mathbf{A}$ должна быть симметричной, а это значит, что исходная и транспонированная матрицы должны быть равны. Для полной корректности этого утверждения поставим в транспонированной матрице на диагональные элементы вместо компонент буквы (x, T^x) компоненты буквы (y, T^y). Тогда условие строгого равенства букв выражается тождеством

$$\mathbf{A} = \mathbf{A}',$$

где $\mathbf{A}'$ — транспонированная матрица $\mathbf{A}$ с заменой диагональных элементов буквы (x, T^x) на диагональные элементы буквы (y, T^y).

Степень сходства букв определяется разностью матриц ($\mathbf{A}, \mathbf{A}'$). Если буквы одинаковы, то эта разность равна нулевой матрице. Если буквы разные, то эта разница существенно отличается от нуля.

Ясно, что недиагональные элементы разности матриц при их знаковом представлении должны быть строго нулевыми, тогда как диагональные элементы разности матриц могут отличаться друг от друга в пределах установленного порога чувствительности dA_{kk} для каждой компоненты. Только в этом случае можно считать буквы одинаковыми. В противном случае буквы считаются различными.

Таким образом, степень сходства и различия букв определяется условиями симметрии и асимметрии матриц и их представляющих. Любопытно также отметить аналогию между матричным представлением букв и матричным представлением операторов. Эта аналогия является конструктивной, так как позволяет формировать матрицы (операторы) слов. Для этого последовательность букв можно рассматривать как последовательное действие операторов, и результат этого действия равен некоторому произведению операторов, которые представляются соответствующим произведением матриц последовательности букв.

Результирующая матрица слова снова может быть представлена в знаковом выражении. В этом случае слово становится некоторой метабуквой. Метабуквы можно сравнивать между собой по принципу симметрии матриц сравнения. Рекурсивно эту технологию можно обобщить на фразы и представить их метасловами с последующей знаковой нормировкой их матриц. Матрицы фраз, сведенных к матрицам метаслов, снова можно сравнивать между собой по условию симметрии их матриц сравнения. В результате сравнения выявлять разнообразие фраз, характеризующих режимы процесса.

Продолжая эту процедуру, мы можем построить словарь метафраз, метаметафраз и т.д. Набор этих вложенных метаметрических (ультраметрических) структур можно продолжать до тех пор, пока не будет исчерпан весь арсенал многообразия языка процесса. Ясно, что такая система является открытой и практически всегда неполной.

Рассмотрим теперь признаки экстремума:

Для максимума:

$$\chi(t_n) \geq \chi(t_{n1}) \text{ и } \chi(t_n) \geq \chi(t_{n2}).$$

Для минимума:

$$\chi(t_n) \leq \chi(t_{n1}) \text{ и } \chi(t_n) \leq \chi(t_{n2}).$$

Здесь $t_{n1} < t_n < t_{n2}$ — последовательность моментов времени, рядом стоящих в точке экстремума t_n.

Для гладких сингулярных процессов этот признак хорошо работает и по нему можно быстро находить точки экстремумов. Однако при наличии высокочастотного шума количество экстремумов резко возрастает и конструктивная процедура для гладких функций становится в этом случае неэффективной. Разрешать возникшие противоречия между конструктивностью (технологичностью) и эффективностью помогают различные технологии фильтрации сигнала, отсекающие высокочастотные составляющие шума.

В настоящее время существует множество способов фильтрации сигнала. Простейшим из них является скользящее суммирование. Для этого случая рассмотрим одну оригинальную процедуру фильтрации. Пусть последовательно рассматриваются суммы S_1, S_2, S_3, одинаковой мощности S. Эти суммы будем рассматривать как операторы, а их наложение друг на друга — как произведение операторов. В наиболее симметричном случае весь интервал, покрываемый оператором и принятый за единицу, выразится так:

$$3S - 2S^2 = 1. \qquad (24)$$

Возможны три характерных режима взаимодействия операторов.

1. Совместность

В этом случае

$$S^2 = S, \quad 3S - 2S = 1, \quad S = 1. \qquad (25)$$

Это значит, что все три суммы наложены друг на друга ($t_{n1} = t_n = t_{n2}$). Это по существу одна экстремальная точка, в которой сравнения вырождаются и конструктивная процедура не работает.

2. Несовместность

В этом случае [18]

$$S^2 = 0, \quad 3S = 1, \quad S = 1/3. \qquad (26)$$

Здесь весь интервал сравнения разбивается на три последовательные непересекающиеся равные части. Минимальное количество точек, реализующих этот режим сравнения, равно трем. Именно три рядом стоящие точки берутся в конструктивной процедуре сравнения в поисках экстремума.

3. Независимость

В этом случае

$$S^2 = S^2, \quad 3S - 2S^2 = 1, \quad (2S - 1)(1 - S) = 0. \quad (27)$$

Новый и содержательный корень операторного уравнения в этом случае есть

$$S = ½, \quad S^2 = ¼. \quad (28)$$

Минимальное количество точек для реализации сравнения в этом случае равно четырем.

И наконец, чтобы отфильтровать высокочастотные вибрации, применим эти два содержательных случая независимо. Тогда минимальное количество точек для реализации такой процедуры равно их произведению

$$n = 3 \cdot 4 = 12 . \quad (29)$$

Итак, имея последовательных двенадцать точек интервала сравнения, разобьем их на три неравные части по четыре точки в каждой. Составим суммы сигналов по каждой четверке. Пусть это будут суммы S_1, S_2, S_3. Тогда максимум определится по условию

$$S_2 \geq S_1 \text{ и } S_2 \geq S , \quad (30)$$

а минимум по сравнению

$$S_2 \leq S_1 \text{ и } S_2 \leq S_3. \quad (31)$$

Затем разобьем интервал сравнения на четыре части по три точки в каждой. Составим сумму сигналов по каждой тройке. Пусть это будут суммы S_1, S_2, S_4. Тогда максимум удовлетворит условию

$$S_3 \geq S_1 \text{ и } S_2 \geq S_4 , \quad (32)$$

а в минимуме будет справедливо выражение

$$S_3 \leq S_1 \text{ и } S_2 \leq S_4 . \quad (33)$$

Решение принимается при одновременном выполнении условий (30, 32) для максимума и (31, 33) для минимума.

Как видим, такая технология сравнений работает на последовательной совокупности трех, четырех и двенадцати точек. Ясно, что значительная часть высокочастотных экстремумов в такой процедуре сравнения будет отфильтрована.

Эту процедуру рекурсивно можно продолжить в сторону укрупнения сумм. Так, если мощность оператора S будет не шесть точек, как в предыдущем случае, а двенадцать (мощность интервала сравнения первой рекурсии), то интервал сравнения второй рекурсии будет содержать 24 точки, которые в первой процедуре сравнения разбиваются на три интервала по 8 точек в каждом, а во второй процедуре сравнения — на 4 интервала по 6 точек в каждом.

Нетрудно убедиться, что n-S рекурсия будет иметь мощность интервала сравнения

$$S^n = 2^n \cdot S^1 = 2^n \cdot 12 . \quad (34)$$

Каждая рекурсия дает более редкое множество точек и более грубое оценивание состояния и режима процесса.

В такой технологии минимальная мощность сравнения равна 12 точкам, следовательно, минимальное число отсчетов в измерении равно 12 на высокочастотном периоде.

ЗАКЛЮЧЕНИЕ

В изложенной выше статье даются некоторые основополагающие результаты нового направления обработки, анализа и интерпретации данных. Четкого термина этого направления еще нет. Однако ряд работ под названием «нестандартный анализ», «конструктивный анализ», «лексикографический анализ», «статистическая фильтрация» явно отражают особенности этого нового направления. В данной работе затронуты основные особенности этих прогрессивных направлений обработки, анализа и интерпретации данных.

Уже первые результаты этих исследований подтвердили возлагаемые на них надежды. Рассмотренные процедуры анализа позволяют совместить, казалось бы, несовместимые требования, а именно вариабельности, стабильности, устойчивости, адаптивности, чувствительности, быстродействия и эффективности.

Многочисленные устройства, основанные на этих процедурах, хорошо зарекомендовали себя как в области распознавания и обработки радио- и акустических сигналов, так и в области медицины (кардиологии).

СПИСОК ЛИТЕРАТУРЫ

1. *Николис Г., Пригожин И.* Познание сложного. Введение. М.: Мир, 1990. 344 с.
2. *Бреббинс К., Уокер С.* Применение метода граничных элементов в технике. М.: Мир, 1982. 248 с.
3. *Флетчер К.* Численные методы на основе метода Галеркина. М.: Мир, 1988. 352 с.
4. Рациональное численное моделирование в нелинейной механике / Сборник под ред. академика О.М. Белоцерковского. М.: Наука, 1990. 224 с.
5. *Трифанов В.* Методические основы синтеза динамических сетей: алгебраическое равновесие и статистика. Л.: препринт ЛИИА, 1981. 31 с.
6. *Стренг Г., Финкс Дж.* Теория метода конечных элементов. М.: Мир, 1977. 349 с.

7. *Хакен Г.* Информация и самоорганизация. Макроскопический подход к сложным системам. М.: Мир, 1991. 161 с.
8. *Жармунский А.В., Кузьмин В.И.* Критические уровни в развитии природных систем. Л.: Наука, 1990. 223 с.
9. *Nesterov M.M., Nesterov V.M., Tarasov N.A.* Simulation of the thin-film growth dynamics and thin-film surface shape. SPb.: SPIIRAS preprint, 1994. 12 p.
10. *Данилов В.Н., Нестеров М.М., Прошин А.П.* К вопросу о построении самосогласованной информационной концепции измерения параметров физических полей // Материалы Первой Международной конференции по проблемам самоорганизации и управления в сложных коммуникационных пространствах, НООТЕХ, СПб., 1997. С.39–41.
11. *Королев О.Ф., Марлей В.Е.* Вычисления в распределенных алгоритмических сетях // Там же. С. 58–60.
12. *Дименберг Ф.М.* Винтовое исчисление и его приложения в механике. М.: Наука, 1965. 199 с.
13. *Цветков В.Д.* Сердце, золотое сечение и симметрия. Пущино: Пущинский научный центр РАН, 1997. 170 с.
14. *Ахапкин Ю.* Биотехника — новое направление компьютеризации. М.: Наука, 1990. 144 с.
15. *Хакен Г.* Синергетика. М.: Мир, 1980. 404 с.
16. *Хакен Г.* Синергетика. Иерархии неустойчивостей в самоорганизующихся системах и устройствах. М.: Мир, 1985. 419 с.
17. *Холодный М. и др.* Методы анализа нелинейных динамических моделей. М.: Мир, 1991. 365 с.

Санкт-Петербургский институт информатики и автоматизации РАН

Материал поступил в редакцию 23.11.99.

UNCONVENTIONAL DATA ANALYSES USING SELF-ORGANIZING TECHNOLOGIES

M. M. Nesterov, V. N. Trifanov, V. N. Danilov

Saint-Petersburg Institute for Informatics and Automation

The paper outlines the basic concepts and general principles of unconventional data analysis using self-organizing analysis technologies and manifestations of hidden organization and periodicity in the hierarhical (ultrametric) mode. Lexographic and constructive data analysis technologies and statistical data filtration as applied to modern instrument engineering are also partially considered.