ПРИМЕНЕНИЕ УЛЬТРАМЕТРИЧЕСКОЙ АДАПТИВНОЙ СТАТИСТИКИ ДЛЯ АНАЛИЗА СТРУКТУРЫ МАСС-СПЕКТРОМЕТРИЧЕСКОГО СИГНАЛА

М. М. Нестеров, В. Н. Данилов, И. Е. Леонов

Санкт-Петербургский институт информатики и автоматизации РАН
199178, Санкт-Петербург, 14-я линия ВО, д. 39
лаборатория ИТЭФИ

УДК 681.3
М. М. Нестеров, В. Н. Данилов, И. Е. Леонов. **Применение ультраметрической адаптивной статистики для анализа структуры масс-спектрометрического сигнала** // Труды СПИИРАН, Вып. 2, т. 2. — СПб.: Наука, 2005.
Аннотация. *Рассмотрен алгоритм создания инвариантных статистических спектров с помощью адаптивных ультраметрических методов моментов и замещающих точек. Построение таких спектров позволит выявлять естественную структуру масс-спектрометрического сигнала.* — Библ. 7 назв.

UDC 681.3
M. M. Nesterov, V. N. Danilov, I. E. Leonov. **Adaptation of ultrametric adaptive statistic for mass-spectrometers signal analysis** // SPIIRAS Proceedings. Issue 2, vol. 2. — SPb.: Nauka, 2005.
Abstract. *adaptation of ultrametric adaptive statistic for mass-spectrometers signal analysisied.*. — Bibl. 7 items.

Введение

В настоящее время потребности экологических и медицинских технологий, в том числе, глобального мониторинга окружающей среды, стимулировали интенсивную разработку малогабаритных времяпролетных масс-спектрометров. При этом, малые размеры приборов и работа в режиме реального времени приводят к необходимости создания и обработки больших массивов информации в наносекундном диапазоне времени. Поэтому актуальными становятся быстрые, устойчивые и эффективные процедуры обработки, анализа и интерпретации данных. Традиционные стандартные технологии статического, динамического и статистического анализ данных являются ресурсоемкими. Как правило, они эффективно работают на массивах малой мощности, описывающих сравнительно гладкие предсказуемые процессы с низким уровнем шума. Однако приведенные выше условия требуют обработки больших и сверхбольших массивов данных со слабой предсказуемостью и высоким уровнем шума.

Особое место в статистических исследованиях занимает проблема стратификации данных, т.е. их разделения на однородные группы, классы, страты, а так же проблема установления границ между соседними стратами и отнесения конкретного наблюдения к той лил иной страте, если оно находится в области границы между ними. Успешное решение этих вопросов дает возможность корректно решать проблемы анализа структуры совокупности организованных данных, в т.ч. структуры масс-спектрометрического сигнала. В решении подобных вопросов теория статистического анализа достигла значительных успехов: в разработке байесовской теории [2, 3]; в методах максимального правдоподобия [4]; в задачах кластеризации по различным критериям сходства и различия, близости и отдаленности данных при их сравнении с разными кластерами и [5] и в других направлениях.

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Однако оставаясь в пределах парадигмы генеральной совокупности с параметрическими законами распределения, традиционная статистика исследовала статистическую организованность данных с точностью до квадратичных форм и статистических моментов второго порядка. Однако при интенсивном развитии систем оценки с точностью до величин второго порядка оказываются нечувствительными и не достаточными. Одним из вариантов решения указанных проблем может быть переход к не традиционной парадигме анализа данных. Новая парадигма основана на отказе от доминирования генеральной совокупности. Принимается другая парадигма, а именно: наблюдаемые данные несут в себе всю информацию об их внутренней организованности.

В рамках парадигмы генеральной совокупности с декларативно выбираемым по воле исследователя параметрическим законом аппроксимации ее распределения, возникает принципиальная трудность связанная со множественностью неопределенностью возможных решений такой аппроксимации по конечной выборке из генеральной совокупности (эта проблема освещена в работе [1]).

Как снять неопределенность возникающую при декларации параметрических законов распределения. Некий просвет в решении упомянутых проблем намечается в более глубоком анализе внутренней организации событий в условии окружающей среды. Трудно согласится с тем что потоки наблюдаемых событий безгранично делимы. Опыт, интуиция и здравый смысл подсказывают, что существуют неделимые кластеры наблюдаемых потоков событий. Наблюдаемые события выделяются в отдельные группы, собирающиеся вокруг точек сгущения. Такая группировка подобна резонансной дифференциации потока событий по резонансным пространственно-временным ритмам.

Многолетний опыт и интуиция позволили исследователям обратить внимание на тот факт [6,7], что состояния системы, наблюдаемые в процессе опыта не распределяются непрерывно в наблюдаемом пространстве S. Они группируются вокруг отдельных точек с состояниями x_k, $k=1,2,\ldots,n$. Это означает, что практически вся информация о состояниях системы находится в области этих узловых точек. В таком случае интегральный закон распределения можно представить в виде

$$F = \int_S f(x)dx = \sum_k a_k f(x_k), \quad k = \overline{1,n}.$$

В этом состоит центральная идея. Континуальная система замещается алгебраической с дискретным множеством состояний x_k с их весами a_k.

Можно записать статистические моменты (первые два) потока событий x_k, наблюдаемых с вероятностью $p_k = \int_{S_k} f(x)dx, \quad x \in S_k$ где S_k – зона тяготения событий x к замещающему состоянию x_k (страта S_k замещающей резонансной точки x_k)

$$J_1 = \sum_k p_k x_k, \quad J_2 = \sum_k p_k {x_k}^2.$$

Порядок статистических моментов, определенных по ограниченной статистической выборке можно продолжить до требуемого значения, чтобы замкнуть систему и получить однозначный закон распределения выборки. В принятых обозначениях эти моменты записываются так

$$\bar{x} = < x >, \quad \overline{x^k} = < (x - \bar{x})^k >, \quad k = \overline{1,n}.$$

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

В системе замещающих точек эти моменты приобретают вид

$$\bar{x} = \sum_i p_i x_i , \quad \overline{x^k} = \sum_i p_i (x_i - \bar{x})^2, \quad i = \overline{1,m} \;\; k = \overline{1,n} .$$

Число и значение замещающих точек, а также вероятность их повторения подлежат определению по наблюдаемой выборке потока событий.

Таким образом, в методе моментов и замещающих точек (методе статистических инвариантов), система замыкается без принятия гипотезы генеральной совокупности в пределах наблюдаемой ограниченной выборки. В рамках этой гипотезы закон распределения наблюдаемого потока событий однозначно восстанавливается. Он по совокупности статистических инвариантов имеет наилучшую настройку в режиме самоорганизации на наблюдаемую выборку. Каждая выборка имеет свой, присущий только ей закон равновесного распределения событий. Закон распределения в этом случае представляет собой паспортную характеристику, отличающую одну конкретную выборку от всех других.

Как правило, традиционные принципы статистического анализа событий не могут вполне удовлетворительно проявлять естественную организацию потока событий. Это осуществляется с той или иной степенью достоверности. Метод статистических инвариантов позволяет повысить достоверность проявления естественной организации потока событий. Причина состоит в том что его (потока) естественная организация в значительной степени управляется статистическими инвариантами, которые можно проявить при статистической обработке данных. Основное свойство статистических инвариантов разных порядков состоит в том, что они не зависят от числа наблюдаемых однородных независимых событий.

Если наблюдаемое событие сложное, и равно сумме элементарных простых событий, то через статистические инварианты сложного события можно выйти на статистические инварианты простых (причинных ненаблюдаемых) событий. В этом случае отклонение от естественного спектра статистических резонансов, в силу отсутствия волевой декларации, сводится к минимум.

Алгоритм построения ультраметрической адаптивной структуры сигнала

Потоки событий организуются в чередующиеся ритмы, которые случайной частотой и периодом. Пусть x — параметр случайного события. Среднее значение суммы независимых событий $\bar{x}_n = < \sum_k x_k > = n\bar{x}, \quad k = \overline{1,n}$ где $\bar{x}$ — среднее значение параметра наблюдаемого потока событий, n — число событий в потоке.

Среднее суммы событий линейно зависит от числа событий в сумме (слово “параметр” для краткости опускаем). Такое свойство суммы Булем называть ее линейным инвариантом. Существует счетное множество линейных инвариантов рассматриваемой суммы. Они выражаются через соответствующие центральные моменты случайной величины.

Примем для удобства упрощенные выражения

$$\overline{x^k} = < (x - \bar{x}) >, \quad k = \overline{1,n} .$$

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$$x_n^k = (\sum_n x)^k .$$

Более подробные математические выкладки приводящихся ниже выражений можно найти в [1]

Для первого, второго и третьего центральных моментов, рассматривая их рекурсивно, можно показать (раскрывая скобки в произведении сумм и учитывая независимость событий)

$$x_n = nx = 0$$

$$x_n^2 = nx^2 = nD, \quad x^2 =< x^2 >= D$$ где D — дисперсия параметра одного события

$$x_n^3 = nx^3$$

Выражения для моментов более высокого порядка оказываются более сложными. Продолжая рассматривать моменты рекурсивно относительно их порядка получим

$$x_n^4 - 3x_n^2 x_n^2 = n(x^4 - 3x^2 x^2)$$

$$x_n^5 - 10x_n^2 x_n^3 = n(x^5 + 10x^2 x^3)$$

$$x_n^6 - 15x_n^2(x_n^4 - 3x_n^2 x_n^2) - 10x_n^3 x_n^3 - 15x_n^2 x_n^2 x_n^2 =$$

$$= n(x^6 - 15x^2(x^4 - 3x^2 x^2) - 10x^3 x^3 - 15x^2 x^2 x^2) .$$

Из приведенного выше видно, что линейные инварианты моментов более высокого порядка рекурсивно выражаются через линейные инварианты более низкого порядка. Запишем получившееся инварианты

$$\overline{x} =< x >;$$

$$x =< (x - \overline{x}) >= 0;$$

$$x_n^2 = D_n = nx^2 = nD;$$

$$J_1 = \frac{x_n^3}{x_n^2} = \frac{x^3}{x^2} = \frac{x_n^3}{D_n} = \frac{x^3}{D};$$

$$J_2 = \frac{x_n^4}{D_n} - 3D_n = \frac{x^4}{D} - 3D \ ;$$

$$J_3 = \frac{x_n^5}{D_n} - 10J_1 D_n = \frac{x^5}{D} - 10J_1 D \ ;$$

$$J_4 = \frac{x_n^6}{D_n} - (15J_2 + 10J_1^2)D_n - 15D_n^2;$$

$$J_4 = \frac{x^6}{D} - (15J_2 + 10J_1^2)D \ - 15D^2;$$

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$$x_m = \min_k x_k \ /$$

$$x_s = \max_k x_k$$

$$S = x_s - x_m$$

Переписав уравнения в виде явного разрешения их относительно центральных моментов

$$x_n^3 = J_1 D_n;$$

$$x_n^4 = J_2 D_n + 3D_n^2;$$

$$x_n^5 = J_3 D_n + 10 J_1 D_n^2;$$

$$x_n^6 = J_4 D_n + (15 J_2 + 10 J_1^2) D_n^2 + 15 D_n^3;$$

$$x^3 = J_1 D \ ;$$

$$x^4 = J_2 D \ + 3D^2;$$

$$x^5 = J_3 D \ + 10 J_1 D^2;$$

$$x^6 = J_4 D \ + (15 J_2 + 10 J_1^2) D^2 + 15 D^3.$$

Из этих выражений видно что центральные моменты, начиная с третьего порядка, разлагаются в ряд по степеням дисперсии события или группы событий. При этом коэффициенты разложения, как функции от инвариантов, сами являются инвариантами. Этот факт позволяет получать функции инвариантов непосредственно путем разложения моментов порядка боле 3 в ряд по дисперсиям. Для этого необходимо рассмотреть соответствующий ряд потока событий с различными количествами слагаемых. По этим слагаемым можно составить соответствующую систему алгебраических уравнений, решением которых и будут соответствующие функции алгебраических инвариантов.

Поскольку инварианты не зависят от числа слагаемых в независимой совокупности событий, они позволяют представить наблюдаемые события как сложные, состоящие из суммы независимых простых событий. По результатам такого представления возникает возможность определения масштабов ненаблюдаемых простых событий как организующего механизма наблюдаемых.

Рассмотрим модель спектрального распределения потока событий с тремя резонансными ритмами $x \in (x_1, x_3, x_2)$ (x_1, x_2 — крайние, x_3 — средний ритм). Обозначим через p_1, p_2, p_3 вероятности возбуждения соответствующих ритмов. Исходная система алгебраических уравнений

$$p_1 + p_2 + p_3 = 1;$$

$$x_1 p_1 + x_2 p_2 + x_3 p_3 = x = 0;$$

$$x_1^2 p_1 + x_2^2 p_2 + x_3^2 p_3 = x^2 = D;$$

$$x_1^3 p_1 + x_2^3 p_2 + x_3^3 p_3 = x^3 = J_1 D;$$

$$x_1^4 p_1 + x_2^4 p_2 + x_3^4 p_3 = x^4 = (J_2 + 3D) D;$$

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$$x_1^5 p_1 + x_2^5 p_2 + x_3^5 p_3 = x^5 = (J_3 + 10J_1 D)D;$$

$$x_1^6 p_1 + x_2^6 p_2 + x_3^6 p_3 = x^6 = (J_4 + (15J_2 + 10J_1^2)D + 15D^2)D.$$

Получая из первых уравнений выражения для вероятностей резонансных ритмов а из последующих возвратную последовательность и вводя S_1, S_2, S_3 как инварианты корней кубического уравнения: $x^3 - S_1 x^2 + S_2 x - S_3 = 0$; можем найти связь дисперсии простого события (D) с инвариантами потока событий (J_k , k=1,4).

Расчетная схема для определения характеристик резонансного спектра потока простых событий с учетом полученных зависимостей имеет вид

$$D^3 - A_1 D^2 + A_2 D - A_3 = 0;$$

$$A_1 = 2(J_1^2 - J_2);$$

$$A_2 = \frac{2(J_4 - J_2^2) + 9(J_2^2 - J_1^4) - 12J_1(J_3 - J_1 J_2)}{12};$$

$$A_3 = \frac{(J_3 - J_1 J_2)^2 - (J_4 - J_2^2)(J_2 - J_1^2)}{12};$$

$$x^3 - S_1 x^2 + S_2 x - S_3 = 0;$$

$$S_1 = \frac{J_3 - J_1 J_2 + 6J_1 D}{J_2 - J_1^2 + 2D};$$

$$-S_2 = J_2 - J_1 S_1 + 3D;$$

$$S_3 = (J_1 - S_1)D.$$

Как видно из уравнения наблюдается фундаментальная связь между дисперсией флуктуации простых событий с инвариантами наблюдаемого потока событий. Тем самым проявляется внутренняя причинная организованность наблюдаемого потока событий. В этом случае проявлен первый внешний уровень причинной организованности наблюдаемого потока событий на резонансной основе внутренних возможностей и внешних ограничений потока. После стратификации всего диапазона ритмов наблюдаемого потока событий по стратам его резонансных ритмов появляется возможность повторить процедуру вскрытия причинных связей для каждой резонансной страты в отдельности. Проявляется второй уровень резонансной организации потока событий. На этом уровне появляются свои страты для которых можно повторить процедуру. Очевидно возникает фрактальная самоорганизация резонансных частот на каждом уровне их бифуркации. Однако на каждом уровне бифуркации спектр ритмов индивидуален и при конечной выборке число уровней конечно. Фрактальный принцип подобия нарушается, поэтому такие структуры будем называть ультраметрическими.

Заключение

Применение описанных выше алгоритмов к масс-спектрометрическому сигналу позволит выявить структуру масс-спектрометрического сигнала наиболее приближенную к естественной организации потока событий.

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Литература

[1] *Трифанов В. Н.* Инвариантный статистический анализ и управление в транспортных системах. — СПб: ”Элмор”, 2003.

[2] *Браунли К. А.* Статистическая теория и методология в науке и технике. — М.: Наука, 1977.

[3] *Феллер В*. Введение в теорию вероятности и ее приложения. — М.: Мир, 1984.

[4] *Элиот Р.* Стохастический анализ и его приложения. — М.: Мир, 1986.

[5] Классификация и кластер. — М.: Мир, 1980.

[6] *Эбелинг В.* Введение в теорию диссипативных структур. — М.: Мир, 1980.

[7] *Хакен Г.* Синергетика. Иерархии неустойчивостей в самоорганизующихся системах и устройствах. — М.: Мир, 1986.