УДК 519.52

ЛОКАЛЬНЫЙ СТАТИСТИЧЕСКИЙ МЕТОД АНАЛИЗА АНСАМБЛЯ ТОЧЕЧНЫХ ОБЪЕКТОВ

В. Г. Жарков[1], Г. Ф. Жарков

Кратко рассмотрен статистический метод выявления локальных закономерностей в распределении конечного числа сложных объектов, обладающих внутренними параметрами. В явном виде описана процедура применения метода к случаю двумерного ансамбля простых точечных объектов.

Одной из задач, часто возникающих при рассмотрении ансамбля локализованных объектов, является выявление возможных закономерностей в их взаиморасположении. Иногда, помимо своих координат, объекты могут быть охарактеризованы и набором внутренних параметров, например, размером, массой, скоростью и т.п. В этом случае возникает дополнительная возможность поиска закономерностей распределения этих параметров по объектам ансамбля. Подобные проблемы встречаются во многих областях естественных наук, в частности, в физике твердого тела (распределение различных дефектов в кристаллах [1]), молекулярной биологии (закономерности появления нуклеотидов и их комбинаций в цепочках ДНК [2]), астрофизике (распределение звездных объектов [3]).

Разработано множество методов, позволяющих решать эти вопросы (см., например, [4]). Практически все они в той или иной мере используют определенное усреднение информации по ансамблю. Так, методы, основанные на Фурье-преобразовании, преобразуют информацию о расположении объектов в прямом x-пространстве в набор параметров, связанных с обратным k-пространством. При этом точное восстановление координат объектов в исходном пространстве на основе имеющихся Фурье-параметров оказывается невозможным. Таким образом, налицо частичная потеря информации при

[1]Институт биологии гена РАН.

подобных преобразованиях. Те же ограничения характерны и для методов, основывающихся на аппарате многочастичных корреляционных функций. В связи с подобным усреднением информации по области, занимаемой ансамблем, можно говорить о глобальном характере имеющихся методов.

Ниже кратко описывается подход [5], позволяющий контролировать степень потери информации при выявлении возможных закономерностей в ограниченных (конечных) ансамблях объектов. В пределе потеря информации может быть сведена к нулю. Можно говорить о локальности предлагаемого статистического подхода, поскольку он анализирует окрестность каждого объекта ансамбля, а затем сравнивает получаемое распределение характеристик этих окрестностей с распределением, получаемым с помощью теории вероятностей (см. ниже).

Основная идея подхода заключается в следующем. Изначально предполагается отсутствие каких-либо предопределенных закономерностей в распределении конечного числа объектов по пространству, занимаемому ансамблем. Другими словами, предполагается, что эти объекты разбросаны случайным образом. Тем не менее, поскольку число объектов и пространство ансамбля конечны, можно ожидать наличия неких локальных флуктуаций "порядка" в распределении объектов для данной реализации ансамбля. В самом деле, возможна ситуация, когда все объекты случайно окажутся в малой подобласти ансамбля. Вероятность этого достаточно мала, но тем не менее подчеркнем, что случайный механизм разбрасывания может создать **любое** распределение объектов. Решение вопроса о том, присутствуют ли какие-либо закономерности в размещении объектов, связан с оценкой вероятности данной реализации ансамбля. Если экспериментальный ансамбль имеет малую вероятность случайной реализации, то есть сильно отклоняется по некоторым параметрам от среднестатистического, то можно с определенной степенью уверенности говорить о наличии неких закономерностей в размещении объектов.

Для оценки вероятности реализации ансамбля предлагается следующая процедура. Ансамбль рассматривается как совокупность окрестностей вокруг каждого объекта с центрами в этих объектах. Размер окрестности достаточно произволен, но часто удобно принимать его равным среднему расстоянию между объектами, как характерному геометрическому параметру (см. [6]). Проводя разбиение окрестности на части (ячейки), можно все более точно описывать распределение соседних объектов, попавших в данную область.

Для количественного определения параметров окрестности вокруг каждого объекта

ансамбля введем понятие конфигурации. Под ней будем понимать совокупность чисел объектов, попавших в ячейки окрестности данного объекта. Например, если число ячеек каждой окрестности равно 3, то совокупность чисел $< n_1, n_2, n_3, n_{out} >$ объектов, находящихся соответственно в первой, второй, третьей ячейках и в остальном объеме ансамбля, и будет представлять конфигурацию вокруг данного объекта. Отметим, что конфигурация может определяться не только распределением координат объектов, но и их внутренних параметров.

Таким образом, исходный ансамбль представляется в виде совокупности конфигураций, число которых равно числу объектов ансамбля. Теория вероятностей позволяет определить 1) вероятность каждой конфигурации при случайном разбрасывании объектов; 2) вероятность реализации того или иного числа одинаковых конфигураций в ансамбле. Если статистические характеристики конфигурации некоторого типа выходят за пределы доверительного интервала, то можно говорить о наличии закономерности в распределении объектов в ансамбле. Поскольку координаты всех конфигураций известны, то известны и координаты объектов, вокруг которых выявлены закономерности этого типа.

В качестве примера приведем простейшую процедуру обработки двумерного ансамбля объектов, не обладающих внутренними характеристиками.

Пусть M – число объектов в ансамбле, занимающем площадь S_0. Окрестность вокруг каждого объекта выберем в виде круга с центром в данном объекте, радиусом r и площадью S_w. Площадь ансамбля вне этой окрестности равна $S_{out} = S_0 - S_w$. Вероятность попадания объекта в данную окрестность равна $p_{in} = S_w/S_0$, а вероятность непопадания равна $p_{out} = S_{out}/S_0$. Ожидаемое число объектов в окрестности равно $m_* = Mp_{in}$.

Если M неразличимых объектов разбрасываются по n различным ячейкам, то вероятность конфигурации $G =< m_1, m_2, ..., m_n >$ равна

$$P = \frac{M!}{m_1!m_2!...m_n!} p^{m_1} p^{m_2} ... p^{m_n}, \qquad (1)$$

где m_i – число частиц в i-ой ячейке, p_i – вероятность частице попасть в i-ую ячейку, $\sum_{i=1}^{n} m_i = M$. Полное число конфигураций равно $V = \frac{(M+n-1)!}{M!(n-1)!}$.

Координату конфигурации примем равной координате выбранного объекта. Разница между имеющимся числом m соседних объектов, попавших в окрестность, и ожидаемым m_* означает присутствие локальной флуктуации плотности в данной окрестности. Оценим вероятность этой флуктуации.

Для нашего случая $n = 2$, $G =< m, M - m - 1 >$ имеем

$$P(m) = \frac{(M-1)!}{m!(M-m-1)!} p^m p^{M-m-1}, \tag{2}$$

поскольку объект в центре конфигурации фиксирован. Величина $P(m)$ представляет вероятность реализации данной конфигурации G.

Для того, чтобы определить статистическую значимость этой конфигурации, необходимо знать число ее реализаций в данном ансамбле. Пусть это число оказалось равным k_0.

Распределение числа реализации k конфигурации G в ансамбле дается выражением

$$W(k) = M! \frac{P^k}{k!} \frac{(1-P)^{M-k}}{(M-k)!}, \tag{3}$$

где $P = P(m)$, $\sum\limits_{k=0}^{M} W(k) = 1$. Зная распределение $W(k)$, задаваясь величиной α доверительной вероятности и пользуясь стандартной логической процедурой математической статистики, можно определить доверительный интервал $C(\alpha)$ значений k:

$$\alpha = \sum_{L}^{R} W(k), \; C(\alpha) = [L, R], \tag{4}$$

где L и R – соответственно левая и правая граница доверительного интервала.

В случае, если экспериментальное значение k_0 числа реализаций данной конфигурации G лежит за пределами доверительного интервала, то можно с вероятностью α утверждать, что конфигурация G представляет собой закономерность в распределении объектов, причем координаты и строение всех k_0 конфигураций типа G известны. Возможность не только выявить наличие закономерностей, но и определить их местонахождения, также отличает предлагаемый подход от метода многочастичных корреляционных функций и методов, основанных на Фурье-разложении.

Обобщение приведенной процедуры на пространства с другим числом измерений, а также на ансамбли из сложных объектов, обладающих внутренними параметрами, не представляет затруднений [6].

На основании изложенного подхода были созданы пакеты программ для компьютерной обработки астрономических данных, распределения дефектов в твердых телах и распределения N-нуклеотидных последовательностей в цепочках ДНК. Результаты применения развитого подхода к конкретным приложениям будут опубликованы отдельно.

Данная работа выполнена при поддержке Российского Фонда фундаментальных исследований, проект 93-04-6450.

ЛИТЕРАТУРА

[1] Жарков В. Г. Диссертация, канд. физ.-мат. наук, ИКАН, М., 1987.
[2] Satchwell S. C. et al. Jour. Mol. Biol., **191**, 659 (1986).
[3] Zharkov G. F. and Zharkov V. G. Astrophys. Journ., in press.
[4] Peebles P. J. E. The Large-Scale of the Universe, Princeton, 1980.
[5] Жарков В. Г., Жарков Г. Ф. Всероссийский авторский патент N 58590 -21-075623, 1991.
[6] Zharkov V. G. and Zharkov G. F. Препринт ФИАН N 82, М., 1991.

Поступила в редакцию 27 октября 1993 г.