УДК 519.234.3+519.234.7

С.И. ЛЯШКО, Д.А. КЛЮШИН, В.В. АЛЕКСЕЕНКО

МНОГОМЕРНОЕ РАНЖИРОВАНИЕ С ПОМОЩЬЮ ЭЛЛИПТИЧЕСКОГО ПИЛИНГА

Ключевые слова: *распознавание образов, диагностика рака, неопределенность, пилинг, доверительные эллипсоиды, дискриминантный анализ, многомерное транжирование.*

ВВЕДЕНИЕ

Системный подход к организму, в частности, к диагностике и лечению рака, является общепринятым. Следует отметить, что важный вклад в развитие этого подхода внес академик В.М. Глушков. В работе «Проблема рака с позиций общей теории систем» [1] он рассмотрел генетические причины рака и сформулировал гипотезу об иммунных методах борьбы со злокачественными клетками. В настоящее время клиницисты и экспериментаторы интенсивно изучают взаимодействие опухоли и организма на цитологическом уровне. Получено множество подтверждений существования опухоль-ассоциированных изменений в здоровых клетках, расположенных далеко от опухоли, в частности, обнаружены и описаны изменения распределения ДНК в клетках буккального эпителия у женщин, больных раком молочной железы [2]. Недавние исследования, выполненные в США [3], непосредственно продемонстрировали возникновение опухоль-ассоциированных повреждений ДНК в клетках, расположенных далеко от опухолей, у мышей, больных саркомой, меланомой и карциномой кишечника. По мнению исследователей, вероятной причиной этих изменений является реакция иммунной системы на злокачественные клетки, а повреждения ДНК вызываются фактором воспаления, вызываемого опухолью. Эти факты хорошо согласуются с гипотезой В.М. Глушкова о генетических и иммунных аспектах рака.

Совместными усилиями сотрудников кафедры вычислительной математики Киевского национального университета имени Тараса Шевченко и Института экспериментальной патологии, онкологии и радиобиологии им. Р.Е. Кавецкого НАН Украины разработаны новые методы распознавания рака с помощью квадратичного дискриминантного анализа, основанного на доверительных эллипсах [4]. Впрочем, несмотря на высокую точность и специфичность предложенного метода, остается нерешенной проблема минимизации неопределенности, возникающей в тех случаях, когда показатели испытуемого пациента попадают в пересечение доверительных эллипсов или не попадают в них вообще. Для решения этой задачи предлагаем применить новый метод многомерного ранжирования с помощью доверительных эллипсов.

МЕТОДЫ МНОГОМЕРНОГО РАНЖИРОВАНИЯ С ПОМОЩЬЮ ПИЛИНГА

Проблема многомерного ранжирования часто возникает при решении многих задач анализа данных, распознавания образов, фильтрации изображений и др. Широкий обзор методов решения этой задачи приведен в работе [5], в которой ранжирование разделено на маргинальное, редуцированное, частичное и условное. Маргинальное ранжирование предусматривает упорядочение выборок по отдельным компонентам. Редуцированное сводится к оценке опреде-

ленного расстояния каждой выборки от некоей центральной точки. Частичное ранжирование распределяет многомерные выборки на группы, внутри которых выборки считаются неразличимыми. Условное ранжирование означает упорядочение многомерных выборок по одному из выбранных компонентов, в то время как остальные элементы ранжируются по правилу, зависящему от выбранного компонента.

Среди основных требований, выдвигаемых к методам многомерного ранжирования данных, следует указать их независимость от предположений о функции распределения и естественный учет ее геометрической природы. Этому условию, в частности, удовлетворяют непараметрические методы редуцированного ранжирования. Одним из наиболее популярных подходов к решению поставленной выше задачи стал подход, основанный на концепции статистической глубины выборок, введенной в работе [5]. Эта концепция позволяет естественным образом ранжировать выборки, автоматически определяя центральные, самые глубокие выборки, а также выбросы, т.е. выборки, находящиеся далеко от центра. На основе этой концепции разработано семейство методов, получивших название «пилинг». Этим методам, в частности, посвящены работы [6–10] и многие другие. Подробную библиографию по этой теме можно найти в [10].

Идея методов пилинга заключается в построении выпуклой фигуры, содержащей заданное множество случайных точек, идентификации точек, лежащих на границе этой фигуры, исключении этих точек из рассмотрения и повторения описанной процедуры для оставшейся части множества. В качестве геометрической основы для перечисленных методов пилинга использовались выпуклая оболочка точек [5], полуплоскость [6], минимальный эллипс [7], симплекс [9]. Преимуществом этих методов является их относительная вычислительная простота, но, к сожалению, все они приводят к неоднозначному ранжированию. В результате возникают группы выборок, имеющих одинаковый ранг и неразличимых с точки зрения их статистической глубины. Количество выборок в таких группах зависит от используемого метода и количества наблюдений. При пилинге с помощью выпуклых оболочек ожидаемое количество точек в таких группах равно количеству вершин и варьируется от 17 для 10^6 точек до 22 для 10^9 точек (по оценкам B. Efron [11] и J. McDermott, D. Lin [12]), а при пилинге с помощью минимальных эллипсов — от одной до пяти в плоскости R^2 и до $(m^2+3m)/2$ в пространстве R^m (по оценкам Б.В. Рублева [13]). Таким образом, остается нерешенной задача построения метода однозначного пилинга, имеющего невысокую вычислительную сложность. Для решения этой задачи предлагается использовать модификацию доверительного эллипсоида, предложенного в работах Ю.И. Петунина, С.И. Ляшко и Б.В. Рублева [13–15], в качестве аппроксимации минимального эллипсоида.

ДОВЕРИТЕЛЬНЫЙ ЭЛЛИПСОИД

Приведем два описания алгоритма: для наглядности подробно опишем построение доверительного эллипса в пространстве R^2, а затем сформулируем общий алгоритм построения доверительного эллипсоида в пространстве R^m при $m>2$. Исходными данными для алгоритма является множество многомерных точек $M_n=\{\vec{x}_1,\ldots,\vec{x}_n\}$.

Случай $m=2$. Построим выпуклую оболочку точек $M_n=\{(x_1, y_1),\ldots,(x_n, y_n)\}$ на плоскости R^2. Найдем две вершины выпуклой оболочки: (x_k, y_k) и (x_l, y_l), расстояние между которыми наибольшее, т.е. точки, лежащие на диаметре выпуклой оболочки. Проведем через точки (x_k, y_k) и (x_l, y_l) прямую L.

Найдем вершины выпуклой оболочки (x_r, y_r) и (x_q, y_q), расстояние которых от прямой L является наибольшим. Проведем через точки (x_r, y_r) и (x_q, y_q) прямые L_1 и L_2, параллельные прямой L. Проведем через точки (x_k, y_k) и (x_l, y_l) две прямые: L_3 и L_4, перпендикулярные к прямой L. Пересечения прямых L_1, L_2, L_3 и L_4 образуют прямоугольник П, стороны которого имеют длины a и b (пусть для определенности $a \le b$). Осуществим поворот и перенос системы координат так, чтобы левый нижний угол прямоугольника был расположен в начале новой системы координат с осями Ox' и Oy', а точки $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ перешли в точки $(x'_1, y'_1), (x'_2, y'_2), \ldots, (x'_n, y'_n)$. Выполним сжатие абсцисс всех точек $(x'_1, y'_1), (x'_2, y'_2), \ldots, (x'_n, y'_n)$ с коэффициентом $\alpha = a / b$ и получим совокупность точек $(\alpha x'_1, y'_1), (\alpha x'_2, y'_2), \ldots, (\alpha x'_n, y'_n)$, которые лежат в квадрате S. Найдем центр (x'_0, y'_0) квадрата S и вычислим расстояния от него до каждой точки $r_1, r_2, \ldots, r_n$. Найдем наибольшее число $R = \max(r_1, r_2, \ldots, r_n)$. Построим круг с центром в точке (x'_0, y'_0) и радиусом R. (Все точки $(\alpha x'_1, y'_1)$, $(\alpha x'_2, y'_2), \ldots, (\alpha x'_n, y'_n)$ теперь лежат внутри этого круга.) Применяя к этому кругу операцию растягивания вдоль оси Ox' с коэффициентом $\beta = 1/\alpha$ и обратные преобразования поворота и переноса, получаем искомый доверительный эллипс. Легко видеть, что средняя сложность этого алгоритма определяется сложностью построения выпуклой оболочки и равна $O(n \lg n)$.

Случай $m > 2$. Построим выпуклую оболочку точек $M_n = \{\vec{x}_1, \ldots, \vec{x}_n\}$ в m-мерном пространстве. Найдем две вершины выпуклой оболочки: $\vec{x}_k$ и $\vec{x}_l$, лежащие на диаметре выпуклой оболочки. Проведем через точки $\vec{x}_k$ и $\vec{x}_l$ прямую L. Осуществим поворот и перенос системы координат так, чтобы диаметр выпуклой оболочки лежал на оси Ox'_1. Строим наименьший прямоугольный параллелепипед, содержащий точки $x'_1, \ldots, x'_n$. Сжимаем пространство, превращая прямоугольный параллелепипед в гиперкуб. Найдем центр $\vec{x}_0$ гиперкуба и вычислим расстояния $r_1, r_2, \ldots, r_n$ от него до каждой точки. Найдем наибольшее число $R = \max(r_1, r_2, \ldots, r_n)$. Построим гипершар с центром в точке $\vec{x}_0$ и радиусом R. Применим к этому гипершару обратные операции растягивания, поворота и переноса, получая требуемый эллипсоид в исходном пространстве. Легко видеть, что средняя сложность этого алгоритма равна $O(n \lg n)$.

Рассмотрим теперь статистические свойства построенного эллипсоида, позволяющие называть его доверительным.

СТАТИСТИЧЕСКИЕ СВОЙСТВА ДОВЕРИТЕЛЬНЫХ ЭЛЛИПСОИДОВ

Пусть $G = (M, F)$, $M \subset R$ — генеральная совокупность случайных величин с неизвестной функцией распределения F. Основной распределенной массой генеральной совокупности G называется подмножество $B \subset M$ такое, что $P\{x \in B\} = 1 - \alpha$, где x — произвольный элемент из выборки, полученной простым случайным выбором из генеральной совокупности G, а α — заданный уровень значимости (например, $\alpha = 0.05$). Если выборочные значения $x_1, x_2, \ldots, x_n$ — симметрично зависимые случайные величины с одинаковой абсолютно непрерывной функцией распределения, то по утверждению Хилла [7]

$$P(x_{n+1} = x \in (x_{(i)}, x_{(j)})) = \frac{j - i}{n + 1},$$

где x_{n+1} — следующее выборочное значение из генеральной совокупности G, а $x_{(i)}, x_{(j)}$ — порядковые статистики. Частичный вариант утверждения Хилла, когда выборочные значения $x_1, x_2, \ldots, x_n$ — независимые в совокупности случайные величинами с одинаковым непрерывным распределением, доказан

в [15]. При некоторых условиях эта формула справедлива, если выборка $(x_1, x_2, \dots, x_n)$ состоит из симметрично зависимых случайных величин с абсолютно непрерывной функцией распределения [16].

Теорема 1. Если $\eta_1, \eta_2, \dots, \eta_{n+1}$ — симметрично зависимые одинаково распределенные случайные величины с абсолютно непрерывной совместной функцией распределения, такие что $P\{\eta_k = \eta_m\} = 0$ при $k \neq m$, то

$$P\{\eta_k \geq \eta_1, \dots, \eta_k \geq \eta_{k-1}, \eta_k \geq \eta_{k+1}, \dots, \eta_k \geq \eta_{n+1}\} = \frac{1}{n+1}.$$

Доказательство. Пусть $F_k(x_1, x_2, \dots x_k)$ — совместная функция распределения любых k величин из $\eta_1, \eta_2, \dots, \eta_{n+1}$. Для произвольного $k = 1, 2, \dots, n+1$ запишем формулу полной вероятности:

$$P\{\eta_k \geq \eta_1, \dots, \eta_k \geq \eta_{k-1}, \eta_k \geq \eta_{k+1}, \dots, \eta_k \geq \eta_{n+1}\} =$$
$$= \int_{R^1} P\{x \geq \eta_1, \dots, x \geq \eta_{k-1}, x \geq \eta_{k+1}, \dots, x \geq \eta_{n+1} / \eta_k = x\} f(x) dx,$$

где $f(x) = F'(x)$ — плотность распределения одной случайной величины.

Заметим, что

$$P\{x \geq \eta_1, \dots, x \geq \eta_{k-1}, x \geq \eta_{k+1}, \dots, x \geq \eta_{n+1} / \eta_k = x\} =$$

$$= \lim_{\varepsilon \to 0} \frac{P\{x \geq \eta_1, \dots, x \geq \eta_{k-1}, x \geq \eta_{k+1}, \dots, x \geq \eta_{n+1}, x - \varepsilon < \eta_k\}}{P\{x - \varepsilon < \eta_k\}} =$$
$$= \lim_{\varepsilon \to 0} \frac{F_{n+1}(x, x, \dots, x, x) - F_{n+1}(x, x, \dots, x, x - \varepsilon)}{F(x) - F(x - \varepsilon)} =$$
$$= \lim_{\varepsilon \to 0} \frac{(F_{n+1}(x, x, \dots, x, x) - F_{n+1}(x, x, \dots, x, x - \varepsilon))/\varepsilon}{(F(x) - F(x - \varepsilon))/\varepsilon} = \frac{\frac{\partial}{\partial y} F_{n+1}(\overbrace{x, \dots, x}^{n}, y)}{f(x)}.$$

Отсюда следует, что

$$P\{\eta_k \geq \eta_1, \dots, \eta_k \geq \eta_{k-1}, \eta_k \geq \eta_{k+1}, \dots, \eta_k \geq \eta_{n+1}\} = \int_{R^1} \frac{\partial}{\partial y} F_{n+1}(x, \dots, x, y) dx.$$

Таким образом, величины $\int_{R^1} \frac{\partial}{\partial y} F_{n+1}(x, \dots x, y)$ одинаковы при $k = 1, 2, \dots, n+1$.

Событие $\{\eta_k \geq \eta_1, \dots, \eta_k \geq \eta_{k-1}, \eta_k \geq \eta_{k+1}, \dots, \eta_k \geq \eta_{n+1}\}$ означает, что некая случайная величина достигнет наибольшего значения среди всех остальных. Эти события образуют полную группу, т.е.

$$\sum_{k=1}^{n+1} P\{\eta_k \geq \eta_1, \dots, \eta_k \geq \eta_{k-1}, \eta_k \geq \eta_{k+1}, \dots, \eta_k \geq \eta_{n+1}\} =$$
$$= (n+1) \int_{R^1} \frac{\partial}{\partial y} F_{n+1}(x, \dots, x, y) dx = 1.$$

Таким образом, $P\{\eta_k \geq \eta_1, \dots, \eta_k \geq \eta_{k-1}, \eta_k \geq \eta_{k+1}, \dots, \eta_k \geq \eta_{n+1}\} = \frac{1}{n+1}$.

Теорема доказана.

Теорема 2. Если $\eta_1, \eta_2, \dots, \eta_{n+1}$ — симметрично зависимые случайные величины с абсолютно непрерывной функцией распределения, такие что $P\{\eta_k = \eta_{(j)}\} = 0$ при $k \neq j$, а $\eta_{(1)} \leq \eta_{(2)} \leq \dots \leq \eta_{(n)}$ — вариационный ряд, построенный по первым n значениям, то $P\{\eta_{n+1} \in (\eta_{(j)}, \eta_{(j+1)}]\} = \frac{1}{n+1}$.

Доказательство. Разделим полуинтервал $(\eta_{(i)}, \eta_{(j+1)}]$ на отрезки:

$$P\{\eta_{n+1} \in (\eta_{(j)}, \eta_{(j+1)}]\} = P\{\eta_{n+1} > \eta_{(j)}, \eta_{n+1} \le \eta_{(j+1)}\} =$$
$$= P\{\eta_{n+1} \in (\eta_{(j)}, \eta_{(j+1)})\} + P\{\eta_{n+1} = \eta_{(j+1)}\} =$$
$$= \sum_{i_1, i_2, \dots, i_n} P\{\eta_{n+1} > \eta_{i_1}, \eta_{n+1} > \eta_{i_1}, \dots, \eta_{n+1} > \eta_{i_j},$$
$$\eta_{n+1} < \eta_{i_{j+1}}, \eta_{n+1} < \eta_{i_{j+2}}, \dots, \eta_{n+1} < \eta_{i_n}\}.$$

Первый член содержит C_n^j слагаемых, т.е. все такие комбинации $\{i_1, i_2, \dots, i_n\}$ чисел $\{1, 2, \dots, n\}$, что j величин среди $\eta_1, \eta_2, \dots, \eta_n$ меньше η_{n+1}, а остальные $n-j$ величины не меньше η_{n+1}. По принципу включения–исключения получаем, что

$$P\{\eta_{n+1} \in (\eta_{(j)}, \eta_{(j+1)}]\} =$$
$$= \sum_{i_1, i_2, \dots, i_n} P\Big\{\underbrace{\eta_{n+1} > \eta_{i_1}, \eta_{n+1} > \eta_{i_1}, \dots, \eta_{n+1} > \eta_{i_j}}_{B},$$
$$\underbrace{\eta_{n+1} \le \eta_{i_{j+1}}}_{A_1}, \underbrace{\eta_{n+1} \le \eta_{i_{j+2}}}_{A_2}, \dots, \underbrace{\eta_{n+1} \le \eta_{i_n}}_{A_{n-j}}\Big\} =$$
$$= \sum_{i_1, i_2, \dots, i_n} \sum_{k=0}^{n-j} (-1)^k \sum_{\substack{\{s_1, s_2, \dots, s_k\} \subseteq \\ \subseteq q\{i_{j+1}, i_{j+2}, \dots, i_n\}}} P\{\eta_{n+1} > \eta_{i_1}, \dots, \eta_{n+1} > \eta_{i_j},$$
$$\eta_{n+1} > \eta_{s_1}, \dots, \eta_{n+1} > \eta_{s_k}\}.$$

Поскольку на интервале $(\eta_{(j)}, \eta_{(j+1)})$ функция распределения F абсолютно непрерывна, по теореме 1 получаем

$$P\{\eta_{n+1} \in (\eta_{(j)}, \eta_{(j+1)}]\} =$$
$$= \sum_{i_1, i_2, \dots, i_n} \sum_{k=0}^{n-j} (-1)^k \sum_{\substack{\{s_1, s_2, \dots, s_k\} \subseteq \\ \subseteq \{i_{j+1}, i_{j+2}, \dots, i_n\}}} \frac{1}{j+k+1} =$$
$$= C_n^j \sum_{k=0}^{n-j} (-1)^k C_{n-j}^k \frac{1}{j+k+1} = C_n^j \sum_{k=0}^{n-j} (-1)^k C_{n-j}^k \int_0^1 t^{j+k} dt =$$
$$= C_n^j \int_0^1 t^j (1-t)^{n-j} dt = C_n^j \int_0^1 \sum_{k=0}^{n-j} (-1)^k C_{n-j}^k t^{j+k} dt.$$

Воспользуемся определением бета-функции:

$$P\{\eta_{n+1} \in (\eta_{(j)}, \eta_{(j+1)}]\} = C_n^j \mathrm{B}(j+1, n-j+1) = C_n^j \frac{\Gamma(j+1)\Gamma(n-j+1)}{\Gamma(n+2)} =$$
$$= C_n^j \frac{j!(n-j)!}{(n+1)!} = \frac{n!}{j!(n-j)!} \frac{j!(n-j)!}{(n+1)!} = \frac{1}{n+1}.$$

Теорема доказана.

Следствие. Из теоремы 2 вытекает, что

$$P\{\eta_{n+1} \in (\eta_{(i)}, \eta_{(j)})\} = \frac{j-i}{n+1} \quad \forall\, 0 \le i < j \le n+1,$$

где $\eta_0 = -\infty, \eta_{n+1} = +\infty$.

Доказательство. Легко видеть, что

$$P\{\eta_{n+1} \in (\eta_{(i)}, \eta_{(j)})\} = P\{\eta_{n+1} \in [\eta_{(i)}, \eta_{(i+1)})\} +$$

$$+ P\{\eta_{n+1} \in [\eta_{(i+1)}, \eta_{(i+2)})\} + \ldots + P\{\eta_{n+1} \in [\eta_{(j-1)}, \eta_{(j)})\} = \frac{j-i}{n+1}.$$

Лемма 1. Если два случайных вектора: $\vec{x}_1 = (\eta_1^1, \eta_2^1, \ldots, \eta_n^1)$ и $\vec{x}_2 = (\eta_1^2, \eta_2^2, \ldots$ $\ldots, \eta_n^2)$, имеют одинаковую функцию плотности распределения f_G, то для любой точки $X_0 = (x_1^0, x_2^0, \ldots, x_n^0)$ расстояния $r_1 = \sqrt{(\eta_1^1 - x_1^0)^2 + (\eta_2^1 - x_2^0)^2 + \ldots + (\eta_n^1 - x_n^0)^2}$ и $r_2 = \sqrt{(\eta_1^2 - x_1^0)^2 + (\eta_2^2 - x_2^0)^2 + \ldots + (\eta_n^2 - x_n^0)^2}$ являются одинаково распределенными случайными величинами.

Доказательство. Построим функцию распределения для случайной величины $r_1 = r(\vec{x}_1)$: $F_{r_1}(x) = P\{r_1 \le x\}$. Для $x < 0$ вероятность равна нулю, так как r_1 — расстояние, величина неотрицательная. Рассмотрим случай $x \ge 0$:

$$P\{r_1 \le x\} = P\left\{\sqrt{(\eta_1^1 - x_1^0)^2 + (\eta_2^1 - x_2^0)^2 + \ldots + (\eta_n^1 - x_n^0)^2} \le x\right\}.$$

Область, для которой $\sqrt{(\eta_1^1 - x_1^0)^2 + (\eta_2^1 - x_2^0)^2 + \ldots + (\eta_n^1 - x_n^0)^2} \le x$, является шаром в n-мерном пространстве с центром в точке X_0 и радиусом x. Обозначим эту область $O^x_{X_0}$. Тогда

$$P\{r_1 \le x\} = P\{\vec{x}_1 \in O^x_{X_0}\} = \iint\limits_{O^x_{X_0}} f_G(t_1, t_2, \ldots, t_n) dt_1 dt_2 \ldots dt_n.$$

Таким образом, показано, что функция распределения r_1 имеет вид

$$F_{r_1}(x) = \begin{cases} 0, & x < 0, \\ \iint\limits_{O^x_{X_0}} f_G(t_1, t_2, \ldots, t_n)\, dt_1 dt_2 \ldots dt_n, & x \ge 0. \end{cases}$$

Поскольку случайные векторы $\vec{x}_1$ и $\vec{x}_2$ имеют одинаковую функцию плотности распределения, то, повторив рассуждения для r_2, получим

$$F_{r_2}(x) = \begin{cases} 0, & x < 0, \\ \iint\limits_{O^x_{X_0}} f_G(t_1, t_2, \ldots, t_n)\, dt_1 dt_2 \ldots dt_n, & x \ge 0. \end{cases}$$

Иначе говоря, $F_{r_1}(x) = F_{r_2}(x)$, это значит, что r_1 и r_2 — одинаково распределенные случайные величины.

Лемма доказана.

Теорема 3. Если векторы $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ являются симметрично зависимыми и одинаково распределенными случайными векторами из генеральной совокупности G, E_n — доверительный эллипсоид, содержащий точки $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$, и $\vec{x}_{n+1} \in G$, то $P(\vec{x}_{n+1} \in E_n) = n/(n+1)$.

Доказательство. После аффинных преобразований функции распределения всех случайных векторов изменяются одинаково. Соответственно $\vec{x}_1', \vec{x}_2', \ldots, \vec{x}_n'$ также являются одинаково распределенными случайными векторами из генеральной совокупности G', которая содержит преобразованные этим же преобразованием точки, в том числе и $\vec{x}_{n+1}'$. Таким образом, точки $\vec{x}_1', \vec{x}_2', \ldots, \vec{x}_n'$ остаются симметрично зависимыми (перестановочными). По лемме 1 случайные величины

$r_1, r_2, \ldots, r_n$ являются одинаково распределенными. Поскольку их порядок зависит только от порядка симметрично зависимых точек $\vec{x}'_1, \vec{x}'_2, \ldots, \vec{x}'_n$, то они также симметрично зависимые. Теперь если для $\vec{x}_{n+1}$ вычислить расстояние r_{n+1} до центра гиперкуба, то по теореме 2

$$P(r_{n+1} < r_{(n)}) = \frac{n}{n+1} = P(\vec{x}_{n+1} \in E_n).$$

Теорема доказана.

Следствие 1. Уровень значимости доверительного эллипсоида, не превышающий 0.05, достигается при $n > 19$.

Следствие 2. Если $n > 19$, то площадь доверительного эллипсоида можно уменьшить без увеличения уровня значимости, удалив наибольшие значения $r_{(l)}, \ldots, r_{(n)}$, при условии, что $l > 19$.

Следствие 3. Поскольку в лемме 1 точка $X_0 = (x_1^0, x_2^0, \ldots, x_n^0)$ произвольная, в качестве отправной точки, от которой вычисляется расстояние до каждой из точек $\vec{x}'_1, \vec{x}'_2, \ldots, \vec{x}'_n$, можно выбирать как центр квадрата, так и центроид множества точек.

Описанному алгоритма свойственно несколько особенностей: 1) на границе эллипсоида всегда лежит только одна точка; 2) эллипсоид содержит n точек с вероятностью $n/(n+1)$, не зависящей от функции распределения; 3) в практических приложениях множество точек всегда можно ограничить 20 выборками, гарантируя уровень значимости, не превышающий 0.05. Это открывает возможности для решения следующих задач.

ВОЗМОЖНОСТИ ДОВЕРИТЕЛЬНЫХ ЭЛЛИПСОИДОВ

Используя доверительные эллипсоиды, можно решить следующие задачи.

Многомерное ранжирование. На первом этапе рассматривается исходное множество точек $M_n = \{\vec{x}_1, \ldots, \vec{x}_n\}$. Строим доверительный эллипсоид E_n, содержащий множество M_n. Исключаем из множества M_n точку $\vec{x}_{(1)}$, лежащую на эллипсоиде E_n. Эта точка имеет наименьшую статистическую глубину. Строим доверительный эллипсоид E_{n-1}, содержащий точки $M_n \setminus \{\vec{x}_{(1)}\}$. Находим точку $\vec{x}_{(2)}$, лежащую на его границе. Продолжая дальше, получим вариационный ряд $\vec{x}_{(1)} < \vec{x}_{(2)} < \ldots < \vec{x}_{(n)}$. Отношение порядка здесь интерпретируется в смысле статистической глубины.

Определение медианы. Точка, имеющая наибольшую статистическую глубину, может рассматриваться как приближение медианы.

Определение выбросов. Точки, имеющие наименьшие значения статистической глубины, являются выбросами.

Определение квантиля заданного уровня. Эллипсоид E_n содержит n точек с вероятностью $n/(n+1)$. Следовательно, для p-го квантиля можно определить значение m такое, что $p \approx n/(n+1)$, и поверхность эллипсоида E_m интерпретировать как аппроксимацию уровня p-го квантиля.

Устранение неопределенности. Допустим, мы хотим разделить с помощью эллипсоидов две обучающих выборки точек: $K_1 = \{x_1, \ldots, x_n\}$ и $K_2 = \{y_1, \ldots, y_m\}$, принадлежащих двум разных классам. Построим доверительные эллипсоиды E_1 и E_2 для множеств K_1 и K_2. Распознавание неизвестной точки z сводится к проверке условий $z \in E_1$ или $z \in E_2$. Если эллипсоиды не пересекаются, то неоднозначность возникает, только если точка не попадает ни в один эллипсоид. Если эллипсоиды пересекаются, то появляется новый вид неоднозначности, когда точка z попадает в пересечение. В этом случае можно определить статистическую глубину точки z в множествах K_1 и K_2 и отнести ее к тому классу, где ее статистическая глубина больше.

ЗАКЛЮЧЕНИЕ

Предложенный метод эллиптического пилинга является эффективным с вычислительной точки зрения и однозначным с точки зрения возникающего порядка выборок. Он допускает точные оценки доверительного уровня эллипсоидов, не зависящие от предположений о функции распределения исходных выборок и их геометрической природы. Этот метод может применяться для снижения неопределенности при квадратичном дискриминантном анализе и при решении других задач распознавания образов, требующих ранжирования многомерных данных (в частности, при кластеризации цифровых изображений).

СПИСОК ЛИТЕРАТУРЫ

1. Глушков В.М. Теория рака с позиций общей теории систем. — Киев, 1979. — 20 с. — (Препр. / ИК АН УССР; 79-26).
2. Цитологическая реактивность онкологического больного / Под ред. К.П. Ганиной. — Киев: Наук. думка, 1995. — 150 с.
3. Redon C.E. et al. Tumors induce complex DNA damage in distant proliferative tissues in vivo // Proc. of the National Academy of Sci. — 2010. — **107**, N 42. — P. 17992–17997.
4. Andrushkiw R.I., Boroday N.V., Klyushin D.A., Petunin Yu.I. Computer-aided cytogenetic method of cancer diagnosis. — New York: Nova Publishers, 2007. — 300 p.
5. Barnett V. The ordering of multivariate data // J. Royal Statist. Soc. Ser. A (General). — 1976. — **139**, N 3. — P. 318–355.
6. Tukey J.W. Mathematics and the picturing of data // Proc. of the Intern. Congress of Mathemat. — Montreal, Canada. — 1975. — P. 523–531.
7. Titterington D.M. Estimation of correlation coefficients by ellipsoidal trimming // Appl. Statist. — 1978. — **27**. — P. 227–234.
8. Oja H. Descriptive statistics for multivariate distributions // Statist. and Probab. Letters. — 1983. — **1**. — P. 327–332.
9. Liu R.J. On a notion of data depth based on random simplices // Annals Statict. — 1990. — **18**. — P. 405–414.
10. Zuo Y., Serfling R. General notions of statistical depth function // Ibid. — 2000. — **28**. — P. 461–482.
11. Efron B. The convex hull of a random set of points // Biometrika. — 1965. — **52**, N 3–4. — P. 331–343.
12. McDermott J.P., Lin D.K.J. Quantile contours and multivariate density estimation for massive datasets via sequential convex hull peeling // IIE Transact. — 2007. — **39**. — P. 581–591.
13. Рубльов Б.В. Квадратичне розпізнавання множин та дослідження гладких метрик. — Автореф. дис. ... докт. фіз.-мат. наук. — Київ, 2004. — 34 с.
14. Петунин Ю.И., Рублев Б.В. Распознавание образов с помощью квадратичных дискриминантных функций // Вычисл. и прикл. математика. — 1996. — Вып. 80. — С. 89–104.
15. Ляшко С.И., Рублев Б.В. Минимальные эллипсоиды и максимальные симплексы в трехмерном евклидовом пространстве // Кибернетика и системный анализ. — 2003. — № 6. — С. 65–70.
16. Мадреимов И., Петунин Ю.И. Характеризация равномерного распределения с помощью порядковых статистик // Теория вероятностей и мат. статистика. — 1982. — **27**. — С. 96–102.
17. Andrushkiw R.I., Klyushin D.A., Petunin Yu.I., Lysyuk V.N. Construction of the bulk of general population in the case of exchangeable sample values // Proc. of the Intern. Conf. of Mathemat. and Engineer. Techniq. in Medicine and Biolog. Sci. (METMBS'03). — Las Vegas, Nevada, USA, June 26–29, 2003. — P. 486–489.

Поступила 21.01.2013