Журнал обчислювальної та прикладної математики **2013, №4(114)** Journal of Computational & Applied Mathematics

УДК 519.71

МНОГОМЕРНОЕ РАНЖИРОВАНИЕ С ПОМОЩЬЮ ЭЛЛИПСОВ ПЕТУНИНА

Д. А. Клюшин, М. В. Присяжная

Резюме. Предложен новый метод ранжирования многомерных выборок с помощью эллипсов Петунина и новая непараметрическая мера близости между многомерными выборками. Показано, при каких условиях асимптотический уровень значимости предложенного критерия не превышает 0,05.

1. Введение

Многие задачи математической морфологии сводятся к ранжированию многомерных выборок. Общепринятая систематизация методов многомерного ранжирования была предложена V. Barnett в работе [1]. В соответствии с этим подходом методы многомерного ранжирования подразделяются на маргинальные, редуцированные, частичные и условные. Маргинальные методы упорядочивают выборки по отдельным компонентам. Редуцированные методы вычисляют расстояние каждой выборки от центра распределения. Частичное ранжирование подразумевает разделение выборок на группы одинаковых выборок. В условных методах производится упорядочение выборок по выбранному компоненту, влияющему на остальные.

В настоящее время широкое распространение среди методов многомерного ранжирования получил подход, основанный на концепции статистической глубины выборок относительно центра распределения и соответствующих методах пилинга. Этим методам, в частности, посвящены работы J. Tukey [2], D. Titterington [3], H. Oja [4], R. Liu [5], Y. Zuo [6], Д. А. Клюшина [7] и других авторов. Эти методы позволяют учитывать геометрические свойства многомерных распределений и являются относительно простыми с вычислительной точки зрения. В этой статье мы предлагаем новый метод упорядочения многомерных данных, основанный на эллипсах Петунина [8]. Отметим, что в отличие от метода, предложенного в работе [7], этот метод не подразумевает пилинга, т.е. выполнения повторяющихся итераций одной и той же процедуры, применяемой к уменьшающемуся множеству точек, а сразу упорядочивает все точки.

2. Эллипсоид Петунина

Не ограничивая общности опишем алгоритм построения эллипсоида Петунина на плоскости, а затем перенесем его в пространство R^m при $m > 2$.

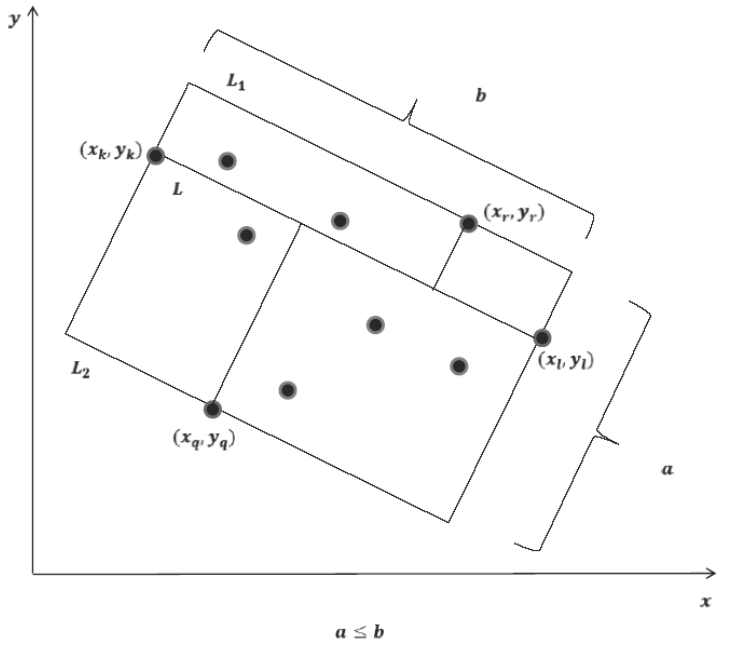

Рис. 1. Прямоугольник Петунина

Исходными данными для алгоритма является множество многомерных точек $M_n = \{\vec{x}_1, ..., \vec{x}_n\}$, где $\vec{x}_n = (x_n, y_n)$.

Эллипс Петунина. На первом этапе построим выпуклую оболочку точек $M_n = \{(x_1, y_1), ..., (x_n, y_n)\}$. Найдем вершины выпуклой оболочки (x_k, y_k) и (x_l, y_l), лежащие на диаметре, т. е. вершины, наиболее удаленные друг от друга. Соединим точки (x_k, y_k) и (x_l, y_l) отрезком L. Найдем вершины выпуклой оболочки (x_r, y_r) и (x_q, y_q), наиболее удаленные от L. Соединим точки (x_r, y_r) и (x_q, y_q) отрезками L_1 и L_2, параллельными отрезку L. Проведем через точки (x_k, y_k) и (x_l, y_l) отрезки L_3 и L_4, перпендикулярные отрезку L. Пересечения отрезков L_1, L_2, L_3 и L_4 образуют прямоугольник П, стороны которого имеют длины a и b (рис. 1).

Будем считать, что $a \le b$. Переведем левый нижний угол прямоугольника в начало новой системы координат с осями Ox' и Oy' с помощью поворота и параллельного переноса. Точки (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) перейдут в точки (x'_1, y'_1), (x'_2, y'_2), ..., (x'_n, y'_n). Отобразим точки (x'_1, y'_1), (x'_2, y'_2), ..., (x'_n, y'_n) в точки $(\alpha x'_1, y'_1)$, $(\alpha x'_2, y'_2)$, ..., $(\alpha x'_n, y'_n)$, где $\alpha = \frac{a}{b}$. В результате получим совокупность точек, лежащих в квадрате S.

Вычислим центр (x'_0, y'_0) квадрата S и найдем расстояния $r_1, r_2, ..., r_n$ от него до каждой точки $(\alpha x'_1, y'_1)$, $(\alpha x'_2, y'_2)$, ..., $(\alpha x'_n, y'_n)$. Наибольшее число $R = \max(r_1, r_2, ..., r_n)$ определяет круг с центром в точке (x'_0, y'_0) и радиусом R. В итоге все точки $(\alpha x'_1, y'_1)$, $(\alpha x'_2, y'_2)$, ..., $(\alpha x'_n, y'_n)$ оказываются

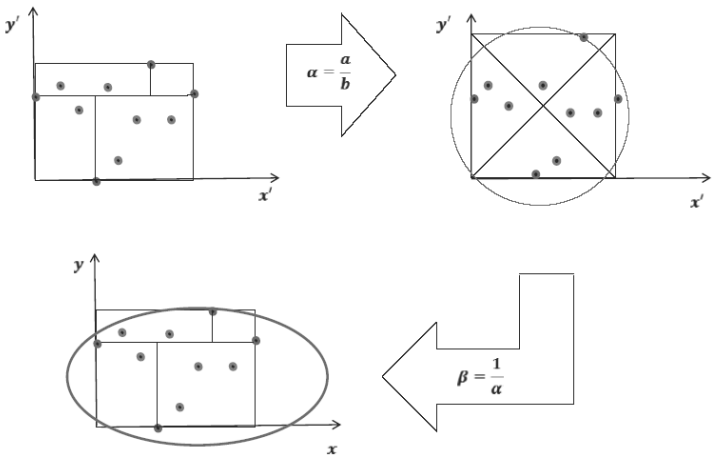

Рис. 2. Построение эллипса Петунина

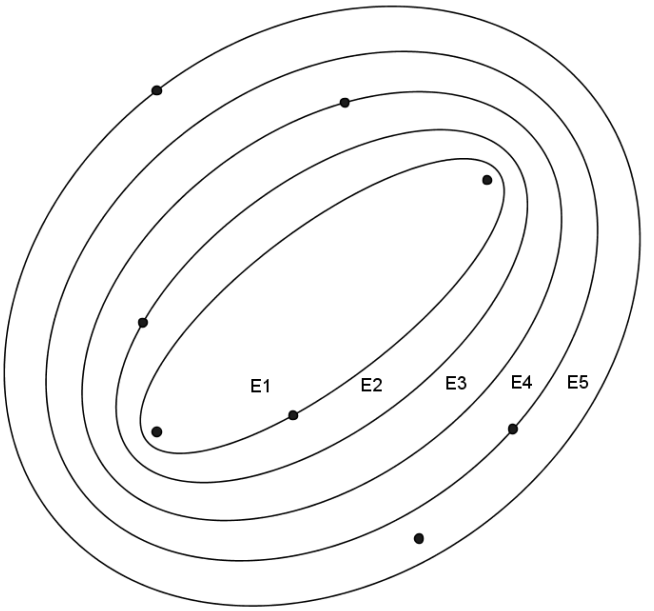

Рис. 3. Вложенные эллипсы Петунина

внутри круга с радиусом R. Растягивая этот круг вдоль оси Ox' с коэффициентом $\beta = \frac{1}{\alpha}$ и выполняя обратные преобразования поворота и переноса, получим эллипс Петунина (рис. 2).

Эллипсоид Петунина. В m-мерном пространстве на первом шаге найдем две вершины выпуклой оболочки $\vec{x}_k$ и $\vec{x}_l$, лежащие на диаметре выпуклой оболочки. Соединим точки $\vec{x}_k$ и $\vec{x}_l$ отрезком L. Повернем и перенесем систему координат, чтобы диаметр выпуклой оболочки лежал на оси Ox'_1. Построим наименьший прямоугольный параллелепипед, содержащий точки $\vec{x}'_1, ..., \vec{x}'_n$.

Сжимая прямоугольный параллелепипед, отобразим точки в гиперкуб. Найдем центр $\vec{x}_0$ гиперкуба и вычислим расстояния $r_1, r_2, ..., r_n$ от него до каждой точки. Найдем наибольшее число $R = \max(r_1, r_2, ..., r_n)$ и построим гипершар с центром в точке $\vec{x}_0$ и радиусом R. Применяя к этому гипершару обратные операции растягивания, поворота и переноса, получим эллипсоид Петунина в m-мерном пространстве.

В результате на каждом вложенном эллипсоиде лежит по одной точке из выборки, т. е.происходит их ранжирование (рис. 3).

Теорема 1 [7]**.** Если векторы $\vec{x}_1, \vec{x}_2, ..., \vec{x}_n$ являются независимыми и одинаково распределенными случайными векторами из генеральной совокупности G, E_n — доверительный эллипсоид, содержащий точки $\vec{x}_1, \vec{x}_2, ..., \vec{x}_n$, и $\overrightarrow{x}_{n+1} \in G$, то $P(\vec{x}_{n+1} \in E_n) = \frac{n}{n+1}$.

3. Новая мера близости между многомерными выборками

По аналогии с p-статистикой, исследованной в статье [9] для одномерного случая, сконструируем меру близости, используя в качестве вариационного ряда многомерные выборки, построенные при ранжировании с помощью эллисоидов Петунина. Вариационному ряду выборок $\vec{x}_{(1)} \preceq \vec{x}_{(2)} \preceq ... \preceq \vec{x}_{(n)}$ поставим в соответствие последовательность вложенных эллипсоидов $E_{(1)} \subset E_{(2)} \subset ... \subset E_{(n)}$. Как следует из теоремы 1, вероятность того, что выборка $\vec{x}$ из многомерной генеральной совокупности G удовлетворяет условию $\vec{x}_{(i)} \preceq \vec{x} \preceq \vec{x}_{(j)}$, равна вероятности попасть между эллипсами $E_{(i)}$ и $E_{(j)}$, т. е. $\frac{j-i}{n+1}$. Это обстоятельство позволяет построить p-статистику для многомерного случая, используя теоремы, доказанные в работе [9].

Напомним основные определения. Пусть $x = (x_1, ..., x_n) \in G$ — выборка из генеральной совокупности G и p — некоторый известный или неизвестный показатель, значения которого могут зависеть от выборки x. Рассмотрим две непрерывные функции $a(u_1, ..., u_n)$ и $b(u_1, ..., u_n)$ от n переменных $u_1, ..., u_n$, удовлетворяющие неравенству $a(u_1, ..., u_n) \le b(u_1, ..., u_n)$ $\forall (u_1, ..., u_n) \in R^n$. Случайный интервал $(a(u_1, ..., u_n), b(u_1, ..., u_n)) = (a, b)$ называется доверительным интервалом для p, соответствующим уровню значимости β, если $P(p \in (a, b)) = 1 - \beta, (0 \le \beta \le 1)$; при этом числа $a = a(x_1, ..., x_n)$, $b = b(x_1, ..., x_n)$ называются доверительными границами для p, соответствующими уровню значимости β[9].

Определение 1 [9]**.** Интервалы $(a_k, b_k) = (a_k(x_1, ..., x_k), b_k(x_1, ..., x_k))$, $k = 1, 2, ...$ называются асимптотическими интервалами для показателей $p_i, i = 1, 2, ..., k, ...$, соответствующими уровню значимости β, если

$$\underline{\lim}_{k \to \infty} P(p_k \in (a_k(x_1, ..., x_k), b_k(x_1, ..., x_k))) = 1 - \beta, \qquad (1)$$

а концы этих интервалов $a_k\,(x_1,...,x_k)$ и $b_k\,(x_1,...,x_k)$ называются асимптотическими доверительными границами.

Определение 2 [9]**.** Величина β называется асимптотическим уровнем значимости последовательности (a_k, b_k), $k = 1, 2, ...$

Определение 3 [9]**.** Если $p_k = p\ \forall k = 1, 2, ...$ то интервал (a_k, b_k) называется асимптотическим доверительным интервалом показателя p, а величина β — асимптотическим уровнем значимости интервала (a_k, b_k).

Обозначим через H гипотезу о равенстве непрерывных функций распределения $F_1(u)$ и $F_2(u)$ генеральных совокупностей многомерных случайных величин G_1 и G_1 соответственно. Пусть $(\vec{x}_1, \vec{x}_2, ..., \vec{x}_n) \in G_1$ и $(\vec{x}'_1, \vec{x}'_2, ..., \vec{x}'_n) \in G_2$, $\vec{x}_{(1)} \preceq \vec{x}_{(2)} \preceq ... \preceq \vec{x}_{(n)}$, $\vec{x}'_{(1)} \preceq \vec{x}'_{(2)} \preceq ... \preceq \vec{x}'_{(n)}$ — соответствующие вариационные ряды. Предположим, что $F_1(u)$=$F_2(u)$. Обозначим через A_{ij}, $k = 1, 2, ..., m$ случайное событие, состоящее в том, что $\vec{x}'_k$ попадает в область $E_{(j)} \backslash E_{(i)}$. Как известно [11], если $F_1(u)$=$F_2(u)$, вероятность p_{ij} этого события вычисляется по формуле:

$$p_{ij}^{(n)} = \frac{j-i}{n+1}. \tag{2}$$

Положим

$$p_{ij}^{(1)} = \frac{h_{ij}m + g^2/2 - g\sqrt{h_{ij}(1-h_{ij})m + g^2/4}}{m+g^2}, \tag{3}$$

$$p_{ij}^{(2)} = \frac{h_{ij}m + g^2/2 + g\sqrt{h_{ij}(1-h_{ij})m + g^2/4}}{m+g^2}, \tag{4}$$

где h_{ij} — частота события A_{ij} в m испытаниях, величина g=3. Рассмотрим доверительные интервалы $I_{ij}^{(n)} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$. Общее количество интервалов $I_{ij}^{(n)}$ равно $N = \frac{n(n-1)}{2}$. Обозначим через L — количество тех интервалов $I_{ij}^{(n)}$, которые содержат вероятности $p_{ij}^{(n)}$. Положим $h = \rho\,(\vec{x}, \vec{x}') = \frac{L}{N}$. Поскольку h — частота случайного события $B = \left\{p_{ij}^{(n)} \in I_{ij}^{(n)}\right\}$, имеющего вероятность $p(B) = 1 - \beta$, то, полагая, $h_{ij} = h, m = N$ и $g = 3$ в формулах (3), (4), получаем доверительный интервал $I^{(n)} = \left(p^{(1)}, p^{(2)}\right)$ для вероятности $p(B)$. Статистика h называется p-статистикой [9]. Она является мерой близости $\rho\,(\vec{x}, \vec{x}')$ между выборками $\vec{x}$ и $\vec{x}'$.

Если гипотеза H истинная, то схема испытаний, в которой могут появляться события $A_{ij}^{(k)}$, называется обобщенной схемой Бернулли, а если гипотеза H неверна — модифицированной схемой Бернулли [10], [11]. В общем случае, когда может быть истинной любая гипотеза, как $F_1(u)$ =$F_2(u)$, так и $F_1(u)$ ≠$F_2(u)$, эта схема испытаний называется МП-схемой [12].

Теорема 2. Если в обобщенной схеме испытаний Бернулли выполняются условия $n = m$, $0 < \lim\limits_{n\to\infty} p_{ij}^{(n)} = p_0 < 1$ и $0 < \lim\limits_{n\to\infty} \frac{i}{n+1} = p^* < 1$, то

асимптотический уровень значимости β последовательности доверительных интервалов $I_{ij}^{(n)}$ для вероятностей $p_{ij}^{(n)}$, построенных по правилу 3s, не превышает 0.05.

Теорема 3. Если выборки $\vec{x} = (\vec{x}_1, ..., \vec{x}_n) \in G_1$ и $\vec{x}' = (\vec{x}'_1, ..., \vec{x}'_m) \in G_2$ имеют одинаковый объем, то асимптотический уровень значимости интервала $I^{(n)} = \left(p^{(1)}, p^{(2)}\right)$, построенный по правилу 3s при $g = 3$ с помощью формул (3), не превосходит 0,05.

Теоремы 2 и 3 являются прямыми следствиями теорем 1 и 2, доказанных в [9], если заменить интервалы $\left(x_{(i)}, x_{(j)}\right)$, образованные порядковыми статистиками одномерной случайной величины x, областями $E_{(j)} \backslash E_{(i)}$, полученными из вариационного ряда $\vec{x}_{(1)} \preceq \vec{x}_{(2)} \preceq ... \preceq \vec{x}_{(n)}$.

Вычислительный эксперимент

В рамках вычислительного эксперимента с помощью статистического пакета R было проведено попарное сравнение выборок из генеральных совокупностей, имеющих двумерное нормальное распределение, векторы математических ожиданий которых равны (0,0), (1,1), (2,2) и (3,3) соответственно, а ковариационная матрица является единичной. Кроме того, было проведено попарное сравнение выборок из генеральных совокупностей, имеющих двумерное нормальное распределение, векторы математических ожиданий которых $(0, 0)$, а на диагонали ковариационной матрицы стоят дисперсии $(1, 1)$, $(2, 2)$, $(3, 3)$ и $(4, 4)$ (внедиагональные элементы равны нулю). Для экспериментов генерировались выборки объемом 300 элементов и вычислена средняя мера близости.

Таблица 1. Усредненная мера близости между выборками

Центры	Мера близости	Дисперсии	Мера близости
(0,0)—(0,0)	0,922	(1,1)—(1,1)	0,922
(0,0)—(1,1)	0,395	(1,1)—(2,2)	0,428
(0,0)—(2,2)	0,120	(1,1)—(3,3)	0,289
(0,0)—(3,3)	0,063	(1,1)—(4,4)	0,223

Как видим, мера близости монотонно убывает по мере увеличения расстояния между центрами распределений при фиксированной дисперсии, а также по мере увеличения дисперсии при фиксированном центре.

Замечание. Предложенная мера близости позволяет проверить гипотезу сдвига и масштаба при одинаковом распределении углов векторов, проведенных из центра распределения к точкам, но, например, если точки двух генеральных совокупностей распределены одинаково, но в противоположных секторах круга, то для их правильного распознавания необходимо учитывать распределение углов векторов.

Выводы

Предложенный метод ранжирования многомерных выборок является точным, поскольку каждая точка выборки получает свой уникальный ранг.

Новая мера близости, основанная на вложенных эллипсах Петунина, позволяет сформулировать критерий для проверки гипотезы о равенстве функций распределения многомерных случайных величин (без учета углового распределения), асимптотический уровень значимости которого не превышает 0,05. Вычислительный эксперимент продемонстрировал практическую ценность разработанной методики.

Література

1. Barnett V. The ordering of multivariate data // Journal of the Royal Statistical Society. Series A (General). — V. 139. — 1976. — №3. — P. 318–355.
2. Tukey J.W. Mathematics and the picturing of data // Proceedings of the International Congress of Mathematician, Montreal, Canada. — 1975. — P. 523–531.
3. Titterington D.M. (1978) Estimation of correlation coefficients by ellipsoidal trimming // Appl. Statist. — 1978. — V. 27. — P. 227–234.
4. Oja H. Descriptive statistics for multivariate distributions // Statistics and Probability Letters. — 1983. — 1. — P. 327–332.
5. Liu R.J. On a notion of data depth based on random simplices //Annals of Statictics. — 1990. — 18. — P. 405–414.
6. Zuo Y., Serfling R. General notions of statistical depth function // Annals of Statistics. — 2000. — 28. — P.461–482.
7. Ляшко С. И., Клюшин Д. А., Алексеенко В. В. Многомерное ранжирование и эллиптический пилинг // Кибернетика и системный анализ. — 2013. — №4. — С. 29—-36.
8. Петунин Ю. И., Рублев Б. В. Распознавание образов с помощью квадратичных дискриминантных функций // Вычисл. и прикл. матем. — 1996. — Вып. 80. — С. 89–104.
9. Клюшин Д. А., Петунин Ю. И. Непараметрический критерий эквивалентности генеральных совокупностей, основанный на мере близости между выборками // Український математичний журнал. — 2003. — Т. 5, №2. — С. 147–163.
10. Матвейчук С. А., Петунин Ю. И. Обобщение схемы Бернулли, возникающее в вариационной статистике. I // Укр. матем. журнал. — 1991. — 42, №4. — С. 518–528.
11. Матвейчук С. А., Петунин Ю. И. Обобщение схемы Бернулли, возникающее в вариационной статистике. II // Укр. матем. журнал. — 1991. — 48, №6. — С. 779–785.
12. Johnson N., Kotz S. Some generalizations of Bernoulli and Polya-Eggenberger contagion models // Statist. Paper. — 1991. — 32. — P.1–17.

Факультет кібернетики, Київський національний університет імені Тараса Шевченка, вул. Володимирська, 64, Київ, 01601, Україна.

Надійшла 15.08.13