Лекція 4. Еліптична функція статистичної глибини даних

Дамо загальне означення функції глибини.

Означення 1. Нехай дано деякий розподіл P у d-вимірному просторі R^d. Довільна функція $D(x;P)$ для кожної $x \in P$, яка буде впорядковувати точки розподілу за спаданням відносно його центру, називається функцією глибини. Значення функції глибини $D(x;P)$ у конкретних точках x розподілу P називається глибиною відповідних точок [9].

Зауваження 1. Центр розподілу визначається в залежності від конкретної функції глибини, яка розглядається (це може бути медіана, центроїд, геометричний центр розподілу тощо).

Фактично функція глибини забезпечує представлення вихідних даних з d-вимірного евклідового простору у вигляді одновимірних. Завдяки цьому можна виділити ряд переваг у використанні функції глибини. Нижче зазначимо найсуттєвіші з них:

1. Єдиність ознаки, за якою проводиться ранжування.
2. Відносна простота ранжування.
3. Незалежність від афінних перетворень і вибору системи координат.

Нижче наведемо класичні властивості функції глибини, введені Y.Zuo і R.Serfling [9], за умови виконання яких, функція глибини називається статистичною функцією глибини у класичному розумінні.

B1. Афінна інваріантність. Глибина точки не залежить від обраної системи координат, а також не змінюється при дії афінного перетворення на вихідні дані.

B2. Максимальність у центрі. Глибина точки, яка є центром розподілу, який розглядається, є найбільшою з-поміж усіх точок розподілу.

B3. Монотонність відносно найглибшої точки. Глибина точок розподілу монотонно спадає відносно направленої уявної прямої, яка проходить через центр розподілу – найглибшу точку розподілу.

B4. Поведінка функції на нескінченності. Якщо $\|x - x_c\| \to \infty$ ($x \in P$ – довільна точка розподілу P, x_c – центр розподілу, $\|\cdot\|$ – звичайна евклідова норма заданого d -вимірного простору R^d), то $D(x;P) \to 0$.

Зауваження 2. Для кращого розуміння практичного застосування конкретних прикладів статистичних функцій глибини надалі будемо розглядати не розподіли в загальному, а вибірки, як скінченні набори даних певного розподілу. Статистичну глибину у даному випадку позначимо, як $D_n(x;M_n)$, де M_n – вибірка, яка розглядається.

Глибина Тьюкі. Щоб розкрити поняття глибини Тьюкі [8] необхідно додатково ввести означення центру вибірки у розумінні Тьюкі.

Означення 2. Нехай дано вибірку M_n деякого розподілу P у d -вимірному просторі R^d . Центром вибірки M_n називається така точка x_c , для якої будь-яка гіперплощина, яка проходить через неї, ділить точки вибірки на дві приблизно рівні підмножини (менша частина повинна мати принаймні $\frac{1}{d+1}$ точок вибірки). Як і медіана, центральна точка не обов'язково повинна бути однією з точок набору даних. Кожний не порожній набір точок (без повторень точок) має принаймні одну центральну точку.

Зауваження 3. У галузі статистики та обчислювальної геометрії, поняття центру (або ще кажуть, центральної точки) є одним з узагальнень поняття медіани в багатовимірному евклідовому просторі.

Означення 3. Нехай дано вибірку M_n деякого розподілу P у d -вимірному просторі R^d . Глибиною Тьюкі точки $x \in M_n$ будемо називати мінімальну кількість точок вибірки, які лежать по один бік від довільної гіперплощини, проведеної через дану точку. Тобто,

$$TD(x_0;P) = \inf\{C(H) \mid C(H), x_0 \in H\},$$

де $C(H)$ – кількість точок, які лежать по один бік від довільної гіперплощини, проведеної через дану точку x_0 .

Лема 1 [8]. Глибина Тьюкі задовільняє умови В1–В4.

Симплексна глибина. Симплексна глибина введена у статті R.J.Liu [6].

Означення 4. Нехай дано вибірку M_n деякого розподілу P у d-вимірному просторі R^d. Симплексною глибиною точки $x \in M_n$ будемо називати кількість симплексів, які побудовані на довільних точках вибірки і які містять точку x.

$$SD_n(x) = \frac{1}{C_n^{d+1}} \sum_{1 \le i_1 \le i_2 \le \ldots \le i_{d+1}} I(x \in S(x_{i_1}, x_{i_2}, \ldots, x_{i_{d+1}})),$$

де $S(x_{i_1}, x_{i_2}, \ldots, x_{i_{d+1}})$ - це симплекс, побудований на $d+1$ точці вибірки M_n, I — індикаторна функція.

Лема 2 [6]. Якщо розподіл, який розглядається, неперервний, то симплексна глибина для нього задовільняє умови В1–В4.

Лема 3 [6]. Якщо розподіл, який розглядається, дискретний, то симплексна глибина для нього може не задовільняти умови В2 та В3.

Глибина Ойя [7] є результатом узагальнення теорії симплексної глибини.

Означення 5. Нехай дано вибірку M_n деякого розподілу P у d-вимірному просторі R^d. Глибиною Ойя точки $x \in M_n$ будемо називати середній об'єм симплексів, які побудовані на довільних d точках вибірки і точці x. Тобто,

$$D_n(x) = \frac{1}{C_n^d} \sum_{1 \le i_1 \le i_2 \le \ldots \le i_d} \nu(x, x_{i_1}, x_{i_2}, \ldots, x_{i_d})),$$

де $\nu(x, x_{i_1}, x_{i_2}, \ldots, x_{i_d})$ — d-вимірний об'єм симплексу у просторі R^d.

Глибина Махаланобіса. Відстань Махалонобіса між точками $x, y \in R^d$ відносно позитивно визначеної матриці M розмірності $d \times d$ обчислюється як

$$d_M^2(x, y) = (x - y)^T M^{-1} (x - y)$$

На основі даної відстані у роботі Y.Zuo і R.Serfling [9] введено поняття глибини Махаланобіса як

$$MHD(x; F) = (1 + \mathrm{d}^2_{\Sigma(F)}(x, \mu(F)))^{-1},$$

де $\mu(F)$ — математичне сподівання розподілу F F; $\Sigma(F)$ — коваріаційна матриця.

Зоноїдальна глибина введена в роботі німецьких математиків G.Koshevoy і G.Mosler [5].

Означення 6. Зоноїдом деякого розподілу P у d-вимірному просторі R^d називається набір точок у просторі R^{d+1}

$$Z = \left\{ z(\mu, h) \mid h : R^d \to [0,1] - \textit{вимірна} \right\}, \text{ де}$$

$$z(\mu, h) = (z_0(\mu, h), \zeta(\mu, h)) \in R^{d+1},$$

$$z_0(\mu, h) = \int_{R^d} h(x) d\mu(x),$$

$$\zeta(\mu, h) = \int_{R^d} x h(x) d\mu(x)$$

Означення 7. Нехай дано вибірку $M_n = \{x_1, ..., x_n\}$ деякого розподілу P у d-вимірному просторі R^d. Зоноїдальною глибиною точки y відносно вибірки M_n називається число $d_z(y \mid x_1, ..., x_n) = \sup\{\alpha : y \in D_\alpha(x_1, ..., x_n)\}$, де

$$D_\alpha(x_1, ..., x_n) = \left\{ \sum_{i=1}^n \lambda_i x_i : \sum_{i=1}^n \lambda_i = 1, 0 \le \lambda_i, \forall i : \alpha\lambda_i \le \frac{1}{n} \right\}$$

Статистична глибина на основі пілінгу опуклими оболонками. Опукла оболонка для набору точок — це мінімальна опукла фігура, що містить задані точки. Іншими словами, це набір точок, які формують "периметр" багатокутника, що утворюється при їх з'єднанні. Пілінг (відсіювання) методом опуклих оболонок досягається систематичним визначенням і видаленням набору точок, які утворюють опуклу оболонку для точок [3].

Нова статистична глибина даних на основі еліпсоїдів Петуніна

У даній роботі ми пропонуємо нове ймовірносне розуміння поняття статистичної функції глибини. Воно у своїй основі має новий метод ранжування, а також ймовірнісну характеристику. Таким чином, крім безпосереднього ранжування елементів

вибірки, ми можемо вирахувати з якою ймовірністю розглядувана точка потрапить у задану вибірку і додатково визначити ранг цієї точки відносно даної вибірки.

Еліпс Петуніна. Для наочності докладно опишемо побудову довірчого еліпса Петуніна в просторі R^2, а потім сформулюємо загальний алгоритм побудови довірчого еліпсоїда в просторі R^m при $m>2$ відповідно до роботи С.І.Ляшка, Д.А. Клюшина та В.В.Алєксєєнко [2]. Вихідними даними для алгоритму є множина точок $M_n = \{\vec{x}_1,...,\vec{x}_n\}$.

Випадок *m*=2. Побудуємо опуклу оболонку точок $M_n = \{(x_1,y_1),...,(x_n,y_n)\}$ на площині R^2. Знайдемо дві вершини опуклої оболонки (x_k,y_k) і (x_l,y_l), відстань між якими є найбільшим, тобто точки, що лежать на діаметрі опуклої оболонки. Проведемо через точки (x_k,y_k) і (x_l,y_l) пряму L. Знайдемо вершини опуклої оболонки (x_r,y_r) і (x_q,y_q), відстань яких від прямої L є найбільшим. Проведемо через точки (x_r,y_r) і (x_q,y_q) прямі L_1 і L_2, паралельні до прямої L. Проведемо через точки (x_k,y_k) і (x_l,y_l) дві прямі L_3 і L_4, перпендикулярні до прямої L. Перетин прямих L_1, L_2, L_3 і L_4 утворять прямокутник Π, сторони якого мають довжини a і b (нехай, для визначеності, $a \le b$). Здійснимо поворот і перенос системи координат, так щоб лівий нижній кут прямокутника був розташований на початку нової системи координат з осями Ox' і Oy', а точки (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) перейшли в точки (x_1',y_1'), (x_2',y_2'), ..., (x_n',y_n'). Виконаємо стискання абсцис усіх точок (x_1',y_1'), (x_2',y_2'), ..., (x_n',y_n') з коефіцієнтом $\alpha = \frac{a}{b}$ і одержимо сукупність точок $(\alpha x_1',y_1')$, $(\alpha x_2',y_2')$, ..., $(\alpha x_n',y_n')$, що лежать у квадраті S. Знайдемо центр (x_0',y_0') квадрата S й обчислимо відстані від нього до кожної точки $r_1,r_2,...,r_n$. Знайдемо найбільше число $R = \max(r_1,r_2,...,r_n)$. Побудуємо коло з центром у точці (x_0',y_0') і радіусом R. (Усі точки $(\alpha x_1',y_1')$,

$(\alpha x_2', y_2')$, ..., $(\alpha x_n', y_n')$ тепер лежать усередині цього кола.) Застосовуємо до цього кола операцію розтягування уздовж осі Ox' з коефіцієнтом $\beta = \frac{1}{\alpha}$ і зворотні перетворення повороту і переносу, одержуючи шуканий довірчий еліпс. Легко бачити, что середня складність цього алгоритму визначається складністю побудови опуклої оболонки і дорівнює $O(n \lg n)$.

Випадок *m*>2. Побудуємо опуклу оболонку точок $M_n = \{\vec{x}_1, ..., \vec{x}_n\}$ у *m*-вимірному просторі. Знайдемо дві вершини опуклої оболонки $\vec{x}_k$ і $\vec{x}_l$, що лежать на діаметрі опуклої оболонки. Проведемо через точки $\vec{x}_k$ і $\vec{x}_l$ пряму L. Здійснимо поворот і перенос системи координат, так щоб діаметр опуклої оболонки лежав на осі Ox_1'. Побудуємо найменший прямокутний паралелепіпед, що містить точки $\vec{x}_1', ..., \vec{x}_n'$. Стиснемо простір, перетворюючи прямокутний паралелепіпед у гіперкуб. Знайдемо центр $\vec{x}_0$ гіперкуба й обчислимо відстані $r_1, r_2, ..., r_n$ від нього до кожної точки. Знайдемо найбільше число $R = \max(r_1, r_2, ..., r_n)$. Побудуємо гіперкулю з центром у точці $\vec{x}_0$ і радіусом R. Застосуємо до цієї гіперкулі зворотні операції розтягування, повороту і переносу, одержуючи необхідний еліпсоїд у вихідному просторі. Легко бачити, що середня складність цього алгоритму дорівнює $O(n \lg n)$.

З демонстраційною метою розглянемо тепер функцію статистичної глибини на площині. Побудуємо довірчі еліпсоїди E_n, на межах яких лежать точки множини $M_n = \{\vec{x}_1, ..., \vec{x}_n\}$. Відповідно до алгоритму Петуніна ці еліпси будуть концентричними, оскільки вони отримані шляхом розтягування концентричних кіл (рис. 1). Ймовірність, що на межі еліпсу Петуніна лежить дві або більше точок, дорівнює нулю. Точки, що лежать на еліпсах E_n, утворюють варіаційний ряд $\vec{x}_{(1)} \prec \vec{x}_{(2)} \prec ... \prec \vec{x}_{(n)}$, де відношення порядку $x \prec y$ означає, що з того, що $x \in E$, $y \in G$ і $x \prec y$ випливає, що $E \supset G$. Інакше кажучи, “найбільша” точка лежить в найбільш “глибокому” еліпсі.

Еліптична статистична глибина (PD). Довірчі еліпси Петуніна дозволяють побудувати нову еліптичну функцію статистичної глибини. Оскільки еліпси Петуніна утворюють лінії рівня цієї функції, тобто точки, що лежать на межах еліпсоїдів E_n мають однакову статистичну глибину, яка дорівнює довірчому рівню еліпса $\frac{n-1}{n+1}$ [2], монотонну і неперервну еліптичну функцію глибини можна побудувати, наприклад, як лінійний сплайн за цими точками. Зрозуміло, що неперервність функції статистичної глибини можна забезпечити не лише лінійними сплайнами.

Лема 5. Еліптична статистична глибина задовільняє умови В1-В4.

Доведення.

В1) Метод побудови еліпсів Петуніна є інваріантним відносно зміни системи координат і афінних перетворень (за рахунок відповідної інваріантності еліпсів і прямокутників).

В2) Якщо багатовимірна випадкова величина має еліптичний розподіл (зокрема, нормальний) то центри еліпсів Петуніна збігається до медіани вибірки, таким чином найбільшого свого значення статистична функція глибини Петуніна набуває саме у цій точці. Для інших розподілів ця функція дає наближення медіани, точність якого залежить від того, наскільки розподіл відрізняється від еліптичного.

В3) $P(x \in E_n) = \frac{n-1}{n+1} \xrightarrow[n\to\infty]{} 1$, тобто при наближенні до центру розподілу, яким у даному випадку є його медіана, ймовірність, а, отже, і значення глибини точки (порядковий номер еліпсу), яка розглядається, монотонно зростає.

В4) Що далі точка x, яку ми розглядаємо, знаходиться від центру даного розподілу, то більшою є ймовірність її потрапляння у еліпс з меншим порядковим номером. Якщо ж $x \to \infty$, то за припущенням компактності генеральної сукупності, яка розглядається, збільшується ймовірність того, що дана точка не потрапить в жодний з еліпсів і таким чином значення її глибини прямує до 0. Доведення закінчено.

Зауваження 5. Легко бачити, що концентричні довірчі еліпси Петуніна із довірчими рівнямиі $\frac{n-1}{n+1}$, визначають новий спосіб упорядкування багатовимірних вибірок.

Глибинно-упорядковані регіони

В роботі I.Cascos [4] наведено означення і основні властивості, які повинні задовольняти глибинно-упорядковані регіони, утворені функцією статистичної глибини.

Означення 8. Глибинно-упорядкованим регіоном (центральним) називається множина точок, статистична глибина яких не меньше наперед заданого значення, тобто

$$D_\alpha(P_X) = \left\{ x \in R^d : D(x; P_X) \geq \alpha \right\},$$

де $D(x; P_X)$ — статистична глибина точки x, взятої із багатовимірної генеральної сукупності із розподілом P_X.

У роботі I.Cascos [4] були введені *бажані* властивості, яким повинні задовільняти глибинно-упорядковані регіони.

1) *Афінна еквіваріантність*: для будь-якого випадкового вектору $x \in R^d$ $D_\alpha(Ax+b) = AD_\alpha(x) + b$ для будь-якої невиродженої матриці A розмірності $d \times d$ і $b \in R^d$.

2) *Вкладеність*: якщо $\alpha \geq \beta$, то $D_\alpha(P_X) \subset D_\beta(P_X)$.

3) *Монотонність*: якщо $x \prec y$ покомпонентно, то $D_\alpha(P_X) + a \subset D_\beta(P_X) + a$, де $a \in R^d$ — довільний додатний вектор.

4) *Компактність*: $D_\alpha(P_X)$ — компактна область.

5) *Опуклість*: $D_\alpha(P_X)$ — опукла область.

6) *Субаддитивність*:

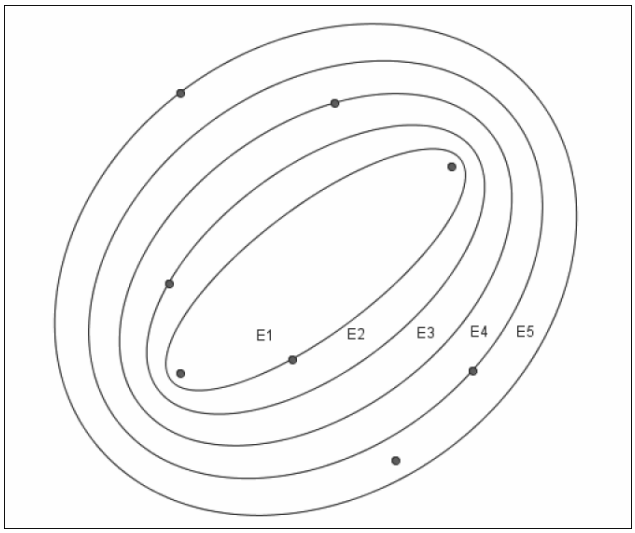

Рис. 1. Еліпси Петуніна

$$D_\alpha\left(P_{X+Y}\right) \subset D_\alpha\left(P_X\right) + D_\alpha\left(P_Y\right)$$

Теорема. Глибинно-упорядковані еліпси Петуніна задовольняють умови 1)–5).

Доведення. Аналіз алгоритму побудови еліпсів Петуніна показує, що властивості 1)–5) (афінна еквіваріантність, вкладеність, монотонність, компактність і опуклість) виконуються очевидним чином.

Cascos [4] зазначив, що субаддитивність — лише *бажана* а не необхідна властивість у дослідженнях фінансових ризиків. Оскільки основним застосуванням еліпсів Петуніна є класифікація багатовимірних даних, а не оцінка фінансових ризиків невиконання властивості субаддитивності глибинно-упорядкованих регіонів не можна вважати її недоліком.

Практичне застосування зазначених вище статистичних функцій глибини

Для порівняння якості і доцільності практичного застосування було проведено розрахунки для всіх зазначених вище статистичних функцій глибини. Параметри: 100 випадків (100 нормально розподілених точок з параметрами (0;1)) та 100 випадків (100 нормально розподілених точок з параметрам (1;1)).

Головним критерієм порівняння доцільності практичного застосування на даному етапі було обрано швидкодію роботи всіх методів.

Основні результати наведені нижче у Таблиці 1.

Функція глибини / Оцінка швидкодії, сек	TD	SD	D	MHD	CHD	PD
(0;1)	5,25	5,05	6,35	6,73	1,36	2,89
(1;1)	5,23	5,20	6,38	6,85	1,38	2,9

Таблиця 1. Порівняння статистичних функцій глибини на основі оцінки швидкодії.

З табл. 1 випливає, що запропонована у даній роботі нова статистична функція глибини (PD) має значні переваги над класичними методами (серед методів точного упорядкування вона забезпечує найвищу швидкодію). Більш висока швидкодія пілінгу опуклими оболонками пояснюється специфікою методу (ранжуються не кожна з точок вибірки окремо, а групи точок).

Список використаних джерел

[1] T.I. Lange, P.F. Mozharovsky Determination of the depth for multivariate data sets // Inductive modelling of complex systems. — 2010. — № 2. — pp. 101–119.(in Russian)

[2] S.I. Lyashko, D.A. Klyushin, V.V. Alexeyenko Mulrivariate ranking using elliptical peeling // Cybernetic and Systems Analysis. — 2013. — № 4. — pp. 29–36.(in Russian)

[3] Barnett V. The ordering of multivariate data // Journal of the Royal Statistical Society. Series A (General).— Vol. 139. — 1976. — No. 3. — pp. 318-355.

[4] Cascos I. Depth function as based of a number of observation of a random vector // Working Paper 07-29. — Statistic and Econometric Series 07. — September 2009, v2. — 28 p.

[5] Koshevoy, G. and Mosler, K. Zonoid trimming for multivariate distributions // Annals of Statistics. — 1997. — 25. — P. 1998-2017.

[6] Liu R.J. On a notion of data depth based on random simplices //Annals of Statictics. — 1990. — 18. —P. 405–414.

[7] Oja H. Descriptive statistics for multivariate distributions // Statistics and Probability Letters. — 1983. — 1. — P. 327–332.

[8] Tukey J. W. Tukey, J.W. Mathematics and the picturing of data // Proceedings of the International Congress of Mathematician, Montreal, Canada. — 1975. — P. 523–531.

[9] Zuo, Y., Serfling, R. General notions of statistical depth function // Ann. Statist. — 2000. — 28. — P. 461-482.