ДОПОВІДІ
НАЦІОНАЛЬНОЇ
АКАДЕМІЇ НАУК
УКРАЇНИ

10 • 2015

ІНФОРМАТИКА ТА КІБЕРНЕТИКА

УДК 519.7

О. А. Галкін

Дослідження непараметричних класифікаторів максимальної глибини на основі просторових квантилів

(Представлено членом-кореспондентом НАН України А. В. Анісімовим)

Запропоновано непараметричний підхід до розв'язання задач розпізнавання, коли розділові поверхні не можуть ефективно апроксимуватися скінченновимірними параметричними лінійними або квадратичними функціями. Підхід грунтується на використанні функції просторової глибини, що є обчислювально дешевшою та може застосовуватися для задач розпізнавання в нескінченновимірних гільбертових просторах. Побудовано глибинний класифікатор на основі концепції просторових квантилів та досліджено його властивості оптимальності у випадку, коли апостеріорні ймовірності конкуруючих еліптичних множин є рівними. Досліджено рівномірну збіжність функції просторової глибини та обчислено оцінки ефективності класифікаторів максимальної глибини.

Ключові слова: байєсівський ризик, просторові квантилі, просторова глибина.

Постановка задачі. Функції глибини даних дозволяють вимірювати центральність r-вимірного елемента даних z відносно заданої r-вимірної множини даних або відносно багатовимірного розподілу Φ. Використовуючи дану концепцію, можна побудувати непараметричний підхід для узагальнення розподільних властивостей одновимірних розподілів до багатовимірних розподілів. Такий підхід може бути застосований до багатовимірної незміщеної статистики, міри багатовимірної дисперсії та асиметрії, а також багатовимірної медіани. Найбільш використовуваними функціями глибини є функція напівпросторової глибини, функція симпліціальної глибини, функція мажоритарної глибини, функція глибини Махаланобіса та функція проекційної глибини.

Функція симпліціальної об'ємної глибини елемента даних z відносно розподілу Φ може бути виражена як

$$\Delta^a(\Phi, z) = \left\{1 + \Omega_\Phi \left[\frac{K\{z, Z_1, \ldots, Z_r\}}{|\Xi_\Phi|^{1/2}}\right]^a\right\}^{-1}, \tag{1}$$

де Ξ_Φ — матриця розсіювання розподілу Φ; $K\{z, Z_1, \ldots, Z_r\}$ — об'єм r-вимірного симплекса, що сформований за допомогою z та $Z_1, \ldots, Z_r$; $Z_1, \ldots, Z_r$ — дані з розподілу Φ. У формулі (1) ділення на $|\Xi_\Phi|^{1/2}$ необхідне для афінно інваріантного перетворення функції глибини.

Застосовуючи концепцію просторових квантилів, введемо поняття глибини розташування, а саме, поняття просторової гибини, що використовується для розвинення методів кластеризації та класифікації [1]. Функція просторової глибини елемента даних z відносно розподілу Φ визначається так:

$$\Delta^{\varphi}(\Phi, z) = 1 - \left\| \Omega_{\Phi}\left\{ \frac{z - Z}{\|z - Z\|} \right\} \right\|, \tag{2}$$

де $Z \approx \Phi$. Величина $\Omega_{\Phi}\{(z - Z)/\|z - Z\|\}$ однозначно визначає функцію розподілу $\Phi(Z)$ та є неперервним та монотонним перетворенням на R^r для всіх Φ при $r \geqslant 2$. У випадку, коли елемент даних z знаходиться недалеко від центра розподілу, $\Omega_{\Phi}\{(z-Z)/\|z - Z\|\}$ буде знаходитися дуже близько до нуля, тому $\Delta^{\varphi}(\Phi, z)$ досягає 1, що є його максимальним значенням. Однак, рухаючись від центра, функція просторової глибини наближатиметься до нуля. На відміну від інших функцій глибини, функцію просторової глибини можна визначити для багатовимірних даних. Крім того, дану функцію можна обчислити для нескінченновимірних гільбертових просторів [2]. Зазначимо, що вибіркові форми функцій глибини можна отримати шляхом заміни Φ на емпіричну функцію розподілу Φ_m, що встановлює масу $1/m$ на кожен з m елементів даних в r-вимірному просторі. Отже, замість $E(\Phi_l, z)$ та $E(\Phi_{m_l}, z)$ використаємо $E(l, z)$ та $E_m(l, z)$ відповідно для позначення теоретичної та емпіричної функції глибини від z в l-й множині даних [3].

Отже, припустимо, що всі розподіли множин даних мають щільності, що є неперервними та додатними в r-вимірному просторі. Класифікатори максимальної глибини не мають довільної параметричної форми розділової поверхні та класифікують елемент даних до класу, відносно якого даний елемент має максимальну глибину розташування. Крім того, такі класифікатори не потребують навчання на вибірці даних, які повинні зберігатися для класифікації нових елементів.

Класифікатори максимальної глибини мажуть бути визначені таким чином:

$$\Im_E(z) = \arg\max_l E_{m_l}(l, z), \tag{3}$$

де $E_{m_l}(l, z)$ — емпірична функція глибини від елемента z в l-й множині даних; m_l — кількість елементів даних у вибірці. Відзначимо, що для розділення конкуруючих множин даних різні концепції глибини можуть бути використані для розробки класифікаторів максимальної глибини, коли всі апріорні ймовірності конкуруючих класів є рівними. Однак, коли конкуруючі множини даних мають однакову матрицю розсіювання, а функція симпліціальної об'ємної глибини використовується для класифікації максимальної глибини, наявність $|\Xi|^{1/2}$ в знаменнику у виразі Δ^a не є обов'язковою. Прикладом цього може бути випадок, коли розподіли множин даних задовольняють модель зсуву розташування [4].

Теорема 1. *Нехай функція щільності $h(z)$ зі сферично-симетричним розподілом є строго спадною на відстані від центра симетрії в розмірності, що є більшою або дорівнює 2. Тоді функція щільності $h(z)$ є функцією просторової глибини.*

Доведення. За точку симетрії приймемо початок координат. Оскільки функція h є інваріантною при ортогональному перетворенні та сферично-симетричною, можна стверджувати, що точки на такій же відстані від центра мають однакову просторову глибину. Далі виберемо такі дві точки z_1 та z_2, що $\|z_1\| < \|z_2\|$, тобто $h(z_1) > h(z_2)$. Отже, на тій же осі координат можна взяти елемент даних в результаті сферичної симетрії [5]. Припустимо, що при $|\kappa_1| < |\kappa_2| z_1 = (\kappa_1, 0, \ldots, 0)$ та $z_2 = (\kappa_2, 0, \ldots, 0)$. Для довільної точки $z^{(1)} = (z_1, z_2, \ldots, z_r)$

можна знайти такі три інші точки $z^{(2)} = (z_1, -z_2, -z_3, \ldots, -z_r)$, $z^{(3)} = (-z_1, z_2, z_3, \ldots, z_r)$ та $z^{(4)} = (-z_1, -z_2, -z_3, \ldots, -z_r)$, що $h(z^{(1)}) = h(z^{(2)}) = h(z^{(3)}) = h(z^{(4)})$. Крім того, вектори вздовж цієї осі координат

$$\sum_{i=1}^{4} \frac{z^{(i)} - z_1}{\|z^{(i)} - z_1\|} h(z^{(i)}) \qquad \text{та} \qquad \sum_{i=1}^{4} \frac{z^{(i)} - z_2}{\|z^{(i)} - z_2\|} h(z^{(i)})$$

направлені до початку координат, де вектор

$$\sum_{i=1}^{4} \frac{z^{(i)} - z_2}{\|z^{(i)} - z_2\|} h(z^{(i)})$$

має більшу розмірність. В результаті, інтегруючи за $z^{(1)}$, $z^{(2)}$, $z^{(3)}$, $z^{(4)}$, отримуємо, що

$$\left\| \Omega_z \left\{ \frac{z_1 - z}{\|z_1 - z\|} \right\} \right\| < \left\| \Omega_z \left\{ \frac{z_2 - z}{\|z_2 - z\|} \right\} \right\|. \tag{4}$$

Теорему доведено.

Далі зазначимо, що

$$|\Psi_{\mathrm{m}} - \Psi| \leqslant \sum_{l=1}^{L} p_l \int \left| \prod_{\substack{i=1 \\ i \neq l}}^{L} \Lambda\{E_{m_l}(l, z) > E_{m_i}(i, z)\} - \prod_{\substack{i=1 \\ i \neq l}}^{L} \Lambda\{E(l, z) > E(i, z)\} \right| h_l(z)\, dz, \tag{5}$$

де h_l — функції щільності класів; $E_{m_l}(l, z)$ — глибина елемента z в l-й множині даних $(l = 1, 2, \ldots, L)$; $E(l, z)$ — множинна глибина елемента z в l-й множині даних $(l = 1, 2, \ldots, L)$; p_l — апріорні ймовірності. Отже, коли класифікатори на основі множинної глибини є оптимальними байєсівськими класифікаторами, результатом буде наслідок теореми Лебега про мажоровану збіжність. Однак це можливо лише у випадку, коли можна показати поточкову збіжність емпіричних функцій глибини до множинних функцій глибини. Зауважимо, що коли еліптичні множини даних відрізняються лише за своїми параметрами розташування, класифікатори на основі множинної глибини є байєсівськими класифікаторами для напівпросторової, симпліціальної, мажоритарної та проекційної глибини [6]. Дане твердження також справедливе для Δ^a при додатковій умові $a \geqslant 1$. Крім того, за умови сферичної симетрії та зсуву розташування просторова функція глибини у множинному вигляді буде байєсівським класифікатором.

Лема 1. *Нехай $h_l(z) = c(z - \varepsilon_l)$ для загальної функції щільності c з $c(kz) \leqslant c(z)$ для кожного z та $k > 1$ та параметра розташування ε_l. Також припустимо, що функції щільності h_1, h_2, …, h_L є еліптично-симетричними. Визначимо Ψ_{m} як частоту помилок емпіричного класифікатора на основі глибини та $\mathrm{m} = (m_1, m_2, \ldots, m_L)$ як вектор розмірів вибірок для різних класів. Частота помилок Ψ_{m} сходиться до оптимального байєсівського ризику при $\min\{m_1, m_2, \ldots, m_L\} \to \infty$ для функції напівпросторової, симпліціальної, мажоритарної та проекційної глибини у випадку рівних апріорних ймовірностей.*

Доведення. У спеціалізованій літературі [7] досліджено рівномірну збіжність емпіричних функцій напівпросторової, симпліціальної, мажоритарної та проекційної глибини. Те ж саме має місце для поточковаої збіжності. Лему доведено.

Деякі функції глибини мають властивість монотонності та є спадними функціями від відстані Махаланобіса, коли розподіл множини даних є еліптичним з функцією щільності, яка є строго спадною в кожному напрямку від її центра симетрії. До даного класу функцій належать такі: функція напівпросторової, симпліціальної, мажоритарної, проекційної глибини, функція симпліціальної об'ємної глибини для $a \geqslant 1$, а також функція глибини Махаланобіса [8]. Тому для класифікації даних згадані функції глибини є еквівалентними відстані Махаланобіса та призводять до оптимального байєсівського класифікатора у випадку рівних апріорних ймовірностей та коли декілька еліптичних множин відрізняються лише своїми параметрами розташування. Функції симпліціальної об'ємної глибини та функції глибини Махаланобіса забезпечують лише лінійну розділову функцію на основі першого та другого моментів даних вибірки. Тому, незважаючи на простоту їх обчислення, вони є досить чутливими до викидів та екстремальних значень. Зазначимо, що більшість класифікаторів на основі функцій глибини не залежать від моментів даних та є більш ефективними, коли дані з вибірки мають розподіли з важкими хвостами.

Лема 2. *Припустимо, що функції щільності $h_1, h_2, \ldots, h_L$ еліптично-симетричні. Визначимо Ψ_{m} як частоту помилок емпіричного класифікатора на основі глибини та $\mathrm{m} = (m_1, m_2, \ldots, m_L)$, що є вектором розмірів вибірок для різних класів. У випадку використання функції просторової глибини частота помилок Ψ_{m} сходиться до оптимального байєсівського ризику при $\min\{m_1, m_2, \ldots, m_L\} \to \infty$, якщо функція c є сферичною.*

Доведення. У спеціалізованій літературі досліджено рівномірну збіжність емпіричних функцій глибини [9]. Результати рівномірної збіжності емпіричної функції просторової глибини до множинної функції просторової глибини досліджено в [10]. Лему доведено.

Відзначимо, що використання функції просторової глибини має значні переваги над іншими функціями глибини. Будучи обчислювально дешевшою, функція просторової глибини може застосовуватися в алгоритмах розв'язання задач класифікації в нескінченновимірних гільбертових просторах. Крім того, оскільки ступінчасті функції напівпросторової глибини, симпліціальної глибини та мажоритарної глибини практично аналогічні, а також враховуючи той факт, що емпірична функція просторової глибини неперервна в z, наявність неоднорідних зв'язків у функціях напівпросторової, симпліціальної та мажоритарної глибини виключена. Також у місцях, де функція щільності зменшується у напрямку від центра, функція просторової глибини має властивість монотонності. Дана властивість має місце у випадку сферично-симетричних розподілів, оскільки функція просторової глибини є інваріантною щодо ортогонального та масштабного перетворення [11].

Лема 3. *Нехай $h_l(z) = c(z - \varepsilon_l)$ для загальної функції щільності c з $(kz) \leqslant c(z)$ для кожного z та $k > 1$, а також параметра розташування ε_l. Також припустимо, що функції щільності h_1, h_2, …, h_L є еліптично-симетричними. Визначимо Ψ_{m} як частоту помилок емпіричного класифікатора на основі глибини та $\mathrm{m} = (m_1, m_2, \ldots, m_L)$ — як вектор розмірів вибірок для різних класів. Для деякого заданого z визначимо $K_l\{z, Z_1, \ldots, Z_r\}$ як об'єм r-вимірного симлекса, сформованого за допомогою z та $Z_1, \ldots, Z_r$, що є елементами даних з h_l. Також припустимо, що $\Omega_{h_l}[K_l\{z, Z_1, \ldots, Z_r\}]^a < \infty$ для всіх $l = 1, 2, \ldots, L$ та деякого $a \geqslant 1$. У випадку симпліціальної об'ємної глибини Δ^a, а також при $\min\{m_1, m_2, \ldots, m_L\} \to \infty$ частота помилок Ψ_{m} сходиться до оптимального байєсівського ризику.*

Доведення. Очевидно, що наявність $|\Xi|^{1/2}$ в знаменнику виразу Δ^a не є обов'язковою, оскільки множини даних задовольняють модель зсуву розташування. Тому при використан-

ні властивостей незміщеної статистики емпірична функція симпліціальної об'ємної глибини Δ^a сходиться майже напевно до множинної функції симпліціальної об'ємної глибини Δ^a для заданого елемента z. Лему доведено.

Цитована література

1. *Jornsten R., Vardi Y., Zhang C. H.* A robust clustering method and visualization tool based on data depth // Statistical data Analysis. – 2002. – P. 354–365.
2. *Chaudhuri P.* On a geometric notion of quantiles for multivariate data // J. of the American Statistical Association. – 1996. – **91**. – P. 864–870.
3. *Zuo Y., Serfling R.* General notions of statistical depth function // The Annals of Statistics. – 2000. – **28**. – P. 463–481.
4. *Lachenbruch P., Mickey M.* Estimation of error rates in discriminant analysis // Technometrics. – 1968. – **10**. – P. 3–10.
5. *Silverman B. W.* Density estimation for Statistics and Data Analysis. – London: Chapman and Hall, 1986. – P. 1–7.
6. *Godtliebsen F., Marron J. S., Chaudhuri P.* Significance in scale space for bivariate density estimation // J. of Computational and Graphical Statistics. – 2002. – **11**. – P. 3–21.
7. *Zuo Y., Serfling R.* Structural properties and convergence results for contours of sample statistical depth functions // The Annals of Statistics. – 2000. – **28**. – P. 484–497.
8. *Nolan D.* Asymptotics for multivariate trimming // Stochastic Processes and Applications. – 1992. – **42**. – P. 159–167.
9. *Koltchinskii V. I.* M-estimation, convexity and quantiles // The Annals of Statistics. – 1997. – **25**. – P. 439–474.
10. *Serfling R.* A depth function and a scale curve based on spatial depth. – Boston: Birkhaeuser, 2002. – P. 27–36.
11. *Holmes C. C., Adams N. M.* A probabilistic nearest neighbor method for statistical pattern recognition // J. of the Royal Statistical Society. – 2002. – **64**. – P. 297–304.

References

1. *Jornsten R., Vardi Y., Zhang C. H.* Statistical data Analysis, 2002: 354–365.
2. *Chaudhuri P.* J. of the American Statistical Association, 1996, **91**: 864–870.
3. *Zuo Y., Serfling R.* The Annals of Statistics, 2000, **28**: 463–481.
4. *Lachenbruch P., Mickey M.* Technometrics, 1968, **10**: 3–10.
5. *Silverman B. W.* Density estimation for Statistics and Data Analysis, London: Chapman and Hall, 1986: 1–7.
6. *Godtliebsen F., Marron J. S., Chaudhuri P.* J. of Computational and Graphical Statistics, 2002, **11**: 3–21.
7. *Zuo Y., Serfling R.* The Annals of Statistics, 2000, **28**: 484–497.
8. *Nolan D.* Stochastic Processes and Applications, 1992, **42**: 159–167.
9. *Koltchinskii V. I.* The Annals of Statistics, 1997, **25**: 439–474.
10. *Serfling R.* A depth function and a scale curve based on spatial depth, Boston: Birkhaeuser, 2002: 27–36.
11. *Holmes C. C., Adams N. M.* J. of the Royal Statistical Society, 2002, **64**: 297–304.

Київський національний університет
ім. Тараса Шевченка

Надійшло до редакції 05.06.2015

А. А. Галкин

Исследование непараметрических классификаторов максимальной глубины на основе пространственных квантилей

Киевский национальный университет им. Тараса Шевченко

Предложен непараметрический подход к решению задач распознавания, когда разделительные поверхности не могут эффективно аппроксимироваться конечномерными параметрическими линейными или квадратичными функциями. Подход основан на использовании функции пространственной глубины, которая является вычислительно дешевле и может применяться для задач распознавания в бесконечномерном гильбертовом пространстве. Построен глубинный классификатор на основе концепции пространственных квантилей, а также исследованы его свойства оптимальности в случае, когда апостериорные вероятности конкурирующих эллиптических множеств равны. Исследована равномерная сходимость функции пространственной глубины, а также рассчитаны оценки эффективности классификаторов максимальной глубины.

Ключевые слова: байесовский риск, пространственные квантили, пространственная глубина.

O. A. Galkin

Research of nonparametric maximum-depth classifiers based on the spatial quantiles

Taras Shevchenko National University of Kiev

A nonparametric approach is proposed to solve the recognition problems, when separating surfaces cannot effectively be approximated by finite-parametric linear or quadratic functions. The approach is based on a function of the spatial depth, which is computationally less expensive and can be used for pattern recognition problems in an infinite-dimensional Hilbert space. A depth-based classifier is built on the basis of the concept of spatial quantiles. The properties of optimality are investigated in the case where the a posteriori probabilities of competing elliptical sets are equal. The uniform convergence of the spatial depth function is studied, and the estimates of the effectiveness of maximum depth classifiers are calculated.

Keywords: Bayes risk, spatial quantiles, spatial depth.