ДОПОВІДІ
НАЦІОНАЛЬНОЇ
АКАДЕМІЇ НАУК
УКРАЇНИ

11 • 2015

УДК 519.7

О. А. Галкін

Асимптотична оцінка глибинних класифікаторів на основі моделі зсуву розташування

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

Досліджується асимптотична поведінка непараметричних класифікаторів симпліціальної, напівпросторової та просторової глибини при відповідних умовах регулярності. Дослідження проводиться для побудови класифікатора максимальної глибини, коли всі апріорні ймовірності конкуруючих класів є рівними та задовольняється модель зсуву розташування. Побудований класифікатор максимальної глибини не залежить від спеціальної параметричної форми розділової поверхні та класифікує елемент даних до класу, відносно якого цей елемент має максимальну глибину розташування. Досліджено випадок нерівних апріорних ймовірностей, коли різні множини даних можуть не належати до спільного сімейства еліптичних розподілів.

Ключові слова: відстань Махаланобіса, байєсівський класифікатор, функція глибини.

Постановка задачі. Незважаючи на значні витрати при обчисленні функцій глибини в задачах великої розмірності, побудова математичного апарату з використанням функцій напівпросторової, симпліціальної та просторової глибини є актуальною задачею у сфері розпізнавання образів. Функція просторової глибини є найбільш простою для обчислення, проте обчислювальні витрати для функцій напівпросторової та симпліціальної глибини швидко збільшуються з розмірністю в геометричній прогресії. В результаті можна використовувати лише наближені модифікації функцій глибини, оскільки точне обчислення даних функцій не є можливим для задач великої розмірності. Застосування наближених модифікацій функцій глибини дозволяє використовувати похідні деякої гладкої функції для визначення напряму найшвидшого підйому або спуску цільової функції, яку необхідно оптимізувати. Даний підхід було застосовано для обчислення функції напівпросторової глибини для задач з розмірністю $r > 2$, а точна модифікація функції напівпросторової глибини використовується лише для двовимірних даних [1].

Починаючи з різних випадкових початкових точок, наближену модифікацію оптимізаційного алгоритму виконано декілька разів для уникнення проблеми можливої наявності

кількох локальних мінімумів. Зауважимо, що оскільки не існує такого наближеного алгоритму для функції симпліціальної глибини, дану функцію глибини було застосовано лише для двовимірних задач.

Досліджуючи емпіричі глибинні класифікатори при відповідних умовах регулярності, наведемо лему щодо асимптотичної точності коефіцієнтів помилкової класифікації.

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 $k > 1$, а також параметра розташування ε_l. Крім того, припустимо, що функції щільності h_1, $h_2, \ldots, h_L$ є еліптично-симетричними. Визначимо Ψ_m як частоту помилок емпіричного класифікатора на основі глибини та $m = (m_1, m_2, \ldots, m_L)$ — як вектор розмірів вибірок для різних класів. Визначимо $E^{0l}(z) = \min\limits_{\{i:\, i \neq l\}} \{E(l, z) - E(i, z)\}$ та Ψ, що дорівнює оптимальному байєсівському ризику. Тоді для деякої функції γ_m, яка залежить від вибору міри глибини, виконується нерівність*

$$\Psi_m < \Psi + \frac{1}{L} \sum_{l=1}^{L} \int\limits_{E^{0l}(z)>0} [1 - \gamma_m\{E^{0l}(z)\}] h_l(z)\, dz, \tag{1}$$

що має місце у випадку рівних апріорних ймовірностей, однак для будь-якого κ ($0 < \kappa < \infty$), $\gamma_m(\kappa) \to 1$ при $\min\{m_1, m_2, \ldots, m_L\} \to \infty$. У даному випадку $\gamma_m(\kappa) = \prod\limits_l \max\{0, 1 - 2e^{-\lfloor m_l/r+1 \rfloor \kappa^2/2}\}$ та $\gamma_m(\kappa) = \prod_l \max\{0, 1 - 2m_l^r e^{-m_l\kappa^2/2}\}$ для функції симпліціальної та напівпросторової глибини відповідно, де $\lfloor z \rfloor$ — найбільше ціле число, що менше або дорівнює z. Якщо розподіли множин даних є сферичними, частота помилок класифікатора на основі функції просторової глибини також задовольняє нерівність (1), *де $\gamma_m(\kappa) = \prod\limits_{l=1}^{L} \max\{0, 1 - 2re^{-m_l\kappa^4/8r^2}\}$.*

Доведення. У випадку, коли $z \in l$-й множині даних, величина Ψ може бути виражена як

$$\Psi = L^{-1} \sum_{l=1}^{L} P\{\arg\max_k E(k, z) \neq l\}, \tag{2}$$

оскільки при заданих умовах класифікатор на основі множинної глибини є оптимальним байєсівським класифікатором [2].

Зауважимо, що вибіркова модифікація функції симпліціальної глибини є незміщеною статистикою з обмеженою функцією ядра. Тому для $l = 1, 2, \ldots, L$ та для кожного $\delta > 0$

$$P\{|E_{m_l}(l, z) - E(l, z)| > \delta\} < 2e^{-2\lfloor m_l/(r+1) \rfloor \delta^2} \tag{3}$$

з використанням нерівності Хефдінга [3]. Отже,

$$L(\Psi_{\mathrm{m}} - \Psi) < \sum_{l=1}^{L} \int\limits_{E^{0l}(z)>0} [1 - \gamma_{\mathrm{m}}^{\circ}\{E^{0l}(z)\}] h_l(z)\, dz \tag{4}$$

для $\gamma_{\mathrm{m}}^{\circ}(\kappa) = \prod\limits_{l=1}^{L} \max\{0, 1 - 2e^{-\lfloor m_l/r+1 \rfloor \kappa^2/2}\}$.

Для функції напівпросторової глибини, де $l = 1, 2, \ldots, L$ та $\delta > 0$, має місце нерівність

$$P\left\{\left|m_l^{-1}\sum_{i=1}^{m_l}\Lambda\{j'(z_{li}-z)>0\} - P\{j'(Z_l-z)>0\}\right| > \delta\right\} < 2e^{-2m_l\delta^2}, \tag{5}$$

що отримана з леми Хефдінга для незалежних та однаково розподілених випадкових величин для будь-якого фіксованого z та j. Множина гіперплощин $\{j'(Z-z)=0\}$ в R^r має розмірність Вапника–Червоненкиса r для деякого фіксованого z [4]. Тому множина вигляду$\{Z\colon j'(Z-z)>0\}$ має поліноміальне розділення, де r — степінь многочлена. Для $l = 1, 2, \ldots, L$ та кожного $\delta > 0$ маємо

$$P\left\{\sup_j\left|m_l^{-1}\sum_{i=1}^{m_l}\Lambda\{j'(z_{li}-z)>0\} - P\{j'(Z_l-z)>0\}\right| > \delta\right\} < 2m_l^r e^{-2m_l\delta^2}, \tag{6}$$

з використанням результатів на ймовірнісних нерівностях на множині $\{Z\colon j'(Z-z)>0\}$. Також зазначимо, що

$$\left|\sup_j m_l^{-1}\sum_{i=1}^{m_l}\Lambda\{j'(z_{li}-z)>0\} - \sup_j P\{j'(Z_l-z)>0\}\right| > \delta. \tag{7}$$

Звідси випливає

$$\sup_j\left|m_l^{-1}\sum_{i=1}^{m_l}\Lambda\{j'(z_{li}-z)>0\} - P\{j'(Z_l-z)>0\}\right| > \delta. \tag{8}$$

В результаті,

$$P\{|E_{m_l}(l,z) - E(l,z)| > \delta\} < 2m_l^r e^{-2m_l\delta^2}. \tag{9}$$

Припустимо, що $E^{01}(z) = \min\limits_{\{l\colon l\neq 1\}}\{E(1,z) - E(l,z)\} > 0$ та виберемо $\delta = E^{01}(z)/2$.

Отже,

$$P\{E_{\mathrm{m}}^{01}(z) > 0\} \geqslant P\{|E_{m_l}(l,z) - E(l,z)| < \frac{E^{01}(z)}{2} \geqslant$$

$$\geqslant \prod_{l=1}^{L}\max\{0, 1 - 2m_l^r e^{-\mathrm{m}_l[E^{01}(z)]^2/2}\} = \gamma_{\mathrm{m}}^*\{E^{01}(z)\} \tag{10}$$

для кожного $l = 1, 2, \ldots, L$.

Оскільки $\gamma_{\mathrm{m}}^*\{E^{01}(z)\} > 0$ та $P\{E_{\mathrm{m}}^{01}(z)\} < 0\} \leqslant 1 - \gamma_{\mathrm{m}}^*\{E^{01}(z)\}$, отримуємо

$$L(\Psi_{\mathrm{m}} - \Psi) = \sum_{l=1}^{L}\int\limits_{E^{0l}(z)>0} P\{E_{\mathrm{m}}^{0l}(z) < 0\}h_l(z)\,dz < \sum_{l=1}^{L}\int\limits_{E^{0l}(z)>0}[1 - \gamma_{\mathrm{m}}^*\{E^{0l}(z)\}]h_l(z)\,dz. \tag{11}$$

У випадку функції просторової глибини визначимо $t_i = (z - z_{li})/\|z - z_{li}\|$ для $i = 1, 2, \ldots, m_l$ та $T = (z - Z)/\|z - Z\|$, де $Z \approx h_l$. Крім того, визначимо $\overline{x}_{m_l} = (1/m_l)\sum\limits_{i=1}^{m_l} t_i$ та $\varepsilon_T = \Omega(T)$. Оскільки $\|\overline{t}_{m_l}\|$ та $\|\varepsilon_T\|$ є додатно-визначеними, маємо

$$P\{|\,\|\overline{t}_{m_l}\| - \|\varepsilon_T\|\,| > \delta\} < \sum_{k=1}^{r} P\left\{|\overline{t}_{m_l}^2(k) - \varepsilon_T^2(k)| > \frac{\delta^2}{r}\right\},$$

де $\overline{t}_{m_l}(k)$ та $\varepsilon_T(k)$ — k-ми компоненти $\overline{t}_{m_l}$ та ε_T відповідно.

Отже, оскільки $|\bar{t}_{m_l}(k) + \varepsilon_T(k)| \leqslant 2$, для кожного $k = 1, 2, \ldots, r$ справедливою є така нерівність:

$$P\left\{|\bar{t}^2_{m_l}(k) - \varepsilon^2_T(k)| > \frac{\delta^2}{r}\right\} < P\left\{|\bar{t}_{m_l}(k) - \varepsilon_T(k)| > \frac{\delta^2}{2r}\right\}. \tag{12}$$

Використовуючи лему Хефдінга, отримуємо нерівність

$$P\left\{|\bar{t}_{m_l}(k) - \varepsilon_T(k)| > \frac{\delta^2}{2r}\right\} < 2e^{-m_l\delta^4/8r^2}, \tag{13}$$

звідки випливає, що

$$P\{|E_{m_l}(l,z) - E(l,z)| > \delta\} = P\{|\,\|\bar{t}_{m_l}\| - \|\varepsilon_T\|\,| > \delta\} = 2re^{-m_l\delta^4/8r^2}, \tag{14}$$

оскільки $\bar{t}_{m_l}(k)$ є середнім значенням незалежних та однаково розподілених обмежених випадкових величин, що знаходиться в інтервалі $[-1, 1]$.

В результаті маємо

$$L(\Psi_{\mathrm{m}} - \Psi) < \sum_{l=1}^{L} \int\limits_{E^{0l}(z)>0} [1 - \gamma^+_{\mathrm{m}}\{E^{0l}(z)\}]h_l(z)\,dz \tag{15}$$

для $\gamma^+_{\mathrm{m}}(\kappa) = \prod\limits_{l=1}^{L} \max\{0, 1 - 2re^{-m_l\kappa^4/8r^2}\}$. Лему доведено.

Відзначимо, що у випадку відповідності моделі зсуву розташування та при рівних апріорних ймовірностях глибинні методи класифікації можуть бути досить ефективними непараметричними та вільними від розподілу методами статистичного аналізу для задач розпізнавання. Однак на практиці різні множини даних можуть не належати до спільного сімейства еліптичних розподілів та мати різні апріорні ймовірності [5]. Далі наведемо лему, яка є теоретичним підгрунтям для побудови глибинних класифікаторів, що дозволяють досягати мінімальних коефіцієнтів помилкової класифікації при умові нерівних апріорних ймовірностей.

Лема 2. *Оптимальний байєсівський класифікатор може бути заданий як*

$$\Im_{\mathrm{opt}}(z) = \arg\max_{l} p_l \lambda_l\{E(l,z)\}, \tag{16}$$

якщо розподіли множин даних є еліптично-симетричними, а для функцій глибини Махаланобіса, напівпросторової, симпліціальної, мажоритарної, симпліціальної об'ємної та проекційної глибини існують деякі функції $\lambda_l(\cdot)$ множинної глибини $E(l,z)$, що залежать від типу функції глибини.

Доведення. Визначимо $C_l = \{(Z_l - \varepsilon_l)'\Xi_l^{-1}(Z_l - \varepsilon_l)\}^{1/2}$, де ε_l — параметр розташування; Ξ_l — матриця розсіювання l-ї множини даних, що має функцію щільності c_l, а $Z_l \approx c_l$.

Відстань Махаланобіса від елемента z з параметром розташування ε_l визначається, як

$$d_l = \{(z - \varepsilon_l)'\Xi_l^{-1}(z - \varepsilon_l)\}^{1/2}. \tag{17}$$

Тому розподіли C_l задаються таким чином:

$$\psi_l(d_l) = \frac{p^{r/2}}{\mathrm{I}r/2}|\Xi_l|^{1/2} d_l^{r-1} c_l(z), \qquad 0 < d_l < \infty, \tag{18}$$

коли функція щільності c_l є еліптично симетричною [6].

Варто зауважити, що відстань Махаланобіса d_l є функцією від множинної глибини $E(l,z)$ у випадку еліптичних множин даних [7]. Крім того, $p_l c_l(z) > p_i c_i(z)$ тоді і тільки тоді, коли $p_l|\Xi_l|^{-1/2}\psi_l(d_l)/d_l^{r-1} > p_i|\Xi_i|^{-1/2}\psi_i(d_i)/d_i^{r-1}$.

В результаті, визначаючи $d_l = \beta_l\{E(l,z)\}$, оптимальний байєсівський класифікатор може бути заданий, як

$$\Im_{\text{opt}}(z) = \arg\max_l p_l \lambda_l\{E(l,z)\}, \tag{19}$$

де $\lambda_l(\kappa) = |\Xi_l|^{-1/2}\psi_l\{\beta_l(\kappa)\}/\{\beta_l(\kappa)\}^{r-1}$. Лему доведено.

Отже, коли функції щільності зменшуються з відстанню Махаланобіса з центра симетрії, а розподіли множин даних задовольняють модель зсуву розташування, функції λ_l є монотонними та однаковими для всіх множин даних. Тому при вищенаведених умовах байєсівський класифікатор (19) є класифікатором максимальної глибини у випадку рівних апріорних ймовірностей.

Цитована література

1. *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.
2. *Serfling R.* A depth function and a scale curve based on spatial depth // Statistics and Data Analysis based on L_1-Norm and Related Methods. – Boston: Birkhäuser, 2002. – P. 27–36.
3. *Hoeffding W.* Probability inequalitites for sums of bounded random variables // J. of the American Statistical Association. – 1963. – **58**. – P. 14–27.
4. *Pollard D.* Convergence of Stochastic Processes. – New York: Springer, 1984. – P. 1–10.
5. *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.
6. *Silverman B. W.* Density estimation for Statistics and Data Analysis. – London: Chapman and Hall, 1986. – P. 1–7.
7. *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.

References

1. *Zuo Y., Serfling R.* The Annals of Statistics, 2000, **28**: 484–497.
2. *Serfling R.* A depth function and a scale curve based on spatial depth. Statistics and Data Analysis based on L_1-Norm and Related Methods, Boston: Birkhäuser, 2002: 27–36.
3. *Hoeffding W.* J. of the American Statistical Association, 1963, **58**: 14–27.
4. *Pollard D.*Convergence of Stochastic Processes, New York: Springer, 1984: 1–10.
5. *Holmes C. C., Adams N. M.* J. of the Royal Statistical Society, 2002, **64**: 297–304.
6. *Silverman B. W.*Density estimation for Statistics and Data Analysis, London: Chapman and Hall, 1986: 1–7.
7. *Jornsten R., Vardi Y., Zhang C. H.* Statistical Data Analysis, 2002: 354–365.

Київський національний університет
ім. Тараса Шевченка, Київ

Надійшло до редакції 22.06.2015

А. А. Галкин

Асимптотическая оценка глубинных классификаторов на основе модели смещения расположения

Киевский национальный университет им. Тараса Шевченко

Исследуется асимптотическое поведение непараметрических классификаторов симплициальной, полупространственной и пространственной глубины при соответствующих условиях регулярности. Исследование проводится для построения классификатора максимальной глубины, когда все априорные вероятности конкурирующих классов равны и удовлетворяется модель смещения расположения. Построенный классификатор максимальной глубины не зависит от специальной параметрической формы разделительной поверхности и классифицирует элемент данных к классу, относительно которого этот элемент имеет максимальную глубину расположения. Исследован случай неравных априорных вероятностей, когда различные множества данных могут не принадлежать общему семейству эллиптических распределений.

Ключевые слова: расстояние Махаланобиса, байесовский классификатор, функция глубины.

O. A. Galkin

Asymptotic estimate of depth-based classifiers within the location shift model

Taras Shevchenko National University of Kiev

The asymptotic behavior of non-parametric simplicial depth, half-space depth, and spatial depth classifiers is studied under appropriate regularity conditions. The research is carried out for the construction of a maximum depth classifier, when all a priori probabilities of all the competing classes are equal, and the location shift model holds. The constructed maximum depth classifier does not depend on the special parametric form of the dividing surface and classifies the data item to a class, with respect to which the element has a maximum depth of location. The case of unequal a priori probabilities is studied, when different data sets may not belong to the common family of elliptical distributions.

Keywords: Mahalanobis distance, Bayesian classifier, depth function.