Доповіді
НАЦІОНАЛЬНОЇ АКАДЕМІЇ НАУК УКРАЇНИ

2 • 2016

ІНФОРМАТИКА ТА КІБЕРНЕТИКА

УДК 519.7 http://dx.doi.org/10.15407/dopovidi2016.02.025

О. А. Галкін

Київський національний університет ім. Тараса Шевченка

E-mail: galkin.a.o.@gmail.com

Афінно-інваріантні глибинні класифікатори на основі методу *k*-найближчих сусідів

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

Досліджуються глибинні класифікатори на основі методу k-найближчих сусідів, що мають непараметричну узгодженність при будь-яких неперервних розподілах. Запропоновано метод симетризації функції глибини, що забезпечує центрально-зовнішнє впорядкування для визначення найближчих сусідів. Побудова симетризації асимптотично гарантує унікальність найглибшої точки, що вирішує проблему опуклої області з нескінченною множиною найглибших точок. Побудованний глибинний класифікатор на основі глибинних околів є афінно-інваріантним, а отже нечутливим до екстремальних значень.

Ключові слова: симетризація, глибинний класифікатор, непараметрична узгодженність.

Постановка задачі. Областю глибини порядкуφ є множина $C_\varphi(P) = \{z \in \mathrm{R}^r : E(z, P) \geqslant \geqslant \varphi\}$ для $\forall\varphi > 0$ та $\forall$ статистичної функції глибини. Оскільки такі області глибини є вкладеними, їх індексація буде відбуватися за ймовірнісною схемою, тобто для $\forall\gamma \in [0, 1)$ величина $R^\gamma(P)$ вказуватиме на найменше значення $R_\varphi(P)$, що має P-ймовірність $\geqslant \gamma$. Тому для глибинних рівнів та ймовірнісного наповнення використовуються верхні та нижні індекси для областей глибини.

Зазначимо, що вибіркові форми верхніх глибин можна отримати шляхом заміни P на відповідний емпіричний розподіл $P^{(m)}$за умови, якщо наявними є r-вимірні дані $Z_1, \ldots, Z_m$. Зазначимо, що Z_i можна впорядкувати таким чином, що

$$E(Z_{(1)}, P^{(m)}) \geqslant E(Z_{(2)}, P^{(m)}) \geqslant \ldots \geqslant E(Z_{(m)}, P^{(m)}), \tag{1}$$

оскільки вибіркова глибина забезпечує центрально-зовнішнє впорядкування елементів даних відносно відповідної найглибшої точки $\overline{\lambda}^{(m)}$. Тому у глибинному сенсі $Z_{(1)}$ є елементом

© О. А. Галкін, 2016

даних, найближчим до $\overline{\lambda}^{(m)}$, $Z_{(2)}$ — другим елементом даних, найближчим до $\overline{\lambda}^{(m)}$, $\ldots$, а $Z_{(m)}$ — елементом даних, найвіддаленішим від $\overline{\lambda}^{(m)}$.

Використовуючи впорядкування (1), статистична функція глибини може застосовуватися для визначення сусідів найглибшої точки $\overline{\lambda}^{(m)}$. Однак для реалізації класифікатора найближчого сусіда необхідно визначити сусідів довільної точки $z \in \mathrm{R}^r$.

У даній роботі пропонується підхід, де глибина розглядається відносно емпіричного розподілу $P_Z^{(m)}$, що пов'язаний з вибіркою, отриманою шляхом додавання до початкових елементів даних Z_1, $Z_2, \ldots, Z_m$ їх відображень $2z - Z_1, \ldots, 2z - Z_m$ відносно z. Зазначимо, що z є асимптотично унікальною найглибшою точкою відносно $P_Z^{(m)}$. Тому відповідна побудова симетризації призводить до z-зовнішнього впорядкування такої форми:

$$E(Z_{Z,(1)}, P_Z^{(m)}) \geqslant E(Z_{Z,(2)}, P_Z^{(m)}) \geqslant \ldots \geqslant E(Z_{Z,(m)}, P_Z^{(m)}), \tag{2}$$

елементи даних якого не є впорядкованими, а використовуються лише для визначення порядку [1].

Отже глибинні околи, тобто вибіркові області глибини $C_{Z,\varphi}^{(m)} = C_\varphi(P_Z^{(m)})$ є критично важливими у даному дослідженні. Зазначимо, що $C_Z^{\gamma(m)}$ означає найменшу область глибини $C_{Z,\varphi}^{(m)}$, що містить щонайменшу частину γ точок Z_1, Z_2, $\ldots$, Z_m. Тому для $\gamma = k/m$, $C_Z^{\gamma(m)}$ є найменшим глибинним околом, що містить k всіх Z_i.

Симетризація функцій глибини. Ми пропонуємо будувати глибинні класифікатори k-найближчих сусідів шляхом заміни евклідових околів на відповідно визначені глибинні околи. Тобто запропонований підхід k-найближчих сусідів буде класифікувати елемент z в множину 1 тоді і тільки тоді, коли множина 1 містить більшу кількість елементів даних, ніж множина 2 в найменшому глибинному околі z, що містить k елементів даних тобто в $C_Z^{\gamma(m)}$, де $\gamma = k/m$.

Отже, запропонований глибинний класифікатор визначається, як

$$\overline{n}_E^{(m)}(z) = \Lambda\left[\sum_{i=1}^{m} \Lambda[X_i = 1] D_i^{\gamma(m)}(z) > \sum_{i=1}^{m} \Lambda[X_i = 0] D_i^{\gamma(m)}(z)\right], \tag{3}$$

де $D_i^{\gamma(m)}(z) = (1/N_z^{\gamma(m)})\Lambda[Z_i \in C_z^{\gamma(m)}]$, а $N_z^{\gamma(m)} = \sum\limits_{l=1}^{m} \Lambda[Z_l \in C_z^{\gamma(m)}]$ визначає число елементів даних в глибинному околі $C_z^{\gamma(m)}$. Зазначимо, що запропонований класифікатор (3) отриманий з використанням глибинної оцінки $\overline{\theta}_E^{(m)}(z)$ умовного математичного сподівання $\theta(z)$, оскільки

$$\overline{n}_E^{(m)}(z) = \Lambda\left[\overline{\theta}_E^{(m)}(z) > \frac{1}{2}\right], \tag{4}$$

де $\overline{\theta}_E^{(m)}(z) = \sum\limits_{i=1}^{m} \Lambda[X_i = 1] D_i^{\gamma(m)}(z)$.

Відзначимо, що в одновимірному випадку $\overline{n}_E^{(m)}$ зводиться до евклідового класифікатора k-найближчих сусідів незалежно від статистичної функції глибини E. Крім того, запропонований класифікатор є афінно-інваріантним, тобто, якщо $Z_1, \ldots, Z_m$ та z підлягають загальній афінній трансформації, результат класифікації залишиться незмінним [2]. У даному випадку спосіб визначення афінно-інваріантного класифікатора k-найближчих сусідів полягає у застосуванні класичного методу k-найближчих сусідів на нормалізованих даних $\overline{\Xi}^{-1/2} Z_i$, $i = 1, \ldots, m$. Для $\forall$ оберненої $r \times r$матриці G та $\forall r$-вектора v, оцінка $\overline{\Xi}$ є афінно-інваріантною оцінкою, що має такий вигляд:

$$\overline{\Xi}(GZ_1 + v, \ldots, GZ_m + v) \propto G\overline{\Xi}(Z_1, \ldots, Z_m)G'. \tag{5}$$

Такий трансформаційний підхід призводить до околів, що не використовують геометрію розподілу в околі точки z та є еліпсоїдами з z-незалежною орієнтацією та формою [3, 4].

Запропонований глибинний класифікатор на основі методу k-найближчих сусідів є узгодженим за відповідних умов. Для цього статистична глибинна функція W повинна задовольняти таким властивостям:

а) властивість унікальної максимізації в центрі симетрії, тобто $E(\lambda, P) > E(z, P)$ для всіх $z \neq \lambda$, якщо P є симетричним відносно λ та має щільність, що є додатньою в λ;

б) властивість узгодженості, тобто величина $\sup\limits_{z \in V} |E(z, P^{(m)}) - E(z, P)| = \mathrm{o}(1)$ майже напевно при $m \to \infty$ для $\forall$ обмеженої r-вимірної борелівської множини V. У даному випадку $P^{(m)}$ є емпіричним розподілом, пов'язаним з m випадковими векторами, що є незалежними однаково розподіленими випадковими величини P;

в) властивість неперервності, тобто $z \mapsto E(z, P)$ є неперервним в околі λ, якщо розподілP є симетричним відносно λ та має щільність, що є додатньою в λ.

Отже, запропоновані класифікатори є узгодженими при практично будь-яких абсолютно неперервних розподілах. Крім того, має місце непараметрична узгодженість, що є відмінною рисою у порівнянні з універсальною узгодженістю стандартного класифікатора k-найближчих сусідів [5].

Теорема 1. *Нехай k_m є послідовністю таких додатніх чисел, що $k_m \to \infty$ та $k_m = o(m)$ при $m \to \infty$. Крім того, нехай E є функцією глибини, що задовольняє властивостям максимальності в центрі, монотонності по відношенню до найглибшої точки, а також властивостях неперервності, унікальної максимізації в центрі симетрії та узгодженості. Припустимо, що $Z|[X = l]$ має функцію щільності h_l, набір точок розриву якої має міру Лебега нуль для $l = 0{,}1$. Тоді, якщо $\sum\limits_m$ є сигма-алгеброю, пов'язаною з (Z_i, X_i), $i = 1, \ldots, m$, має місце узгодженість глибинного класифікатора k_m-найближчих сусідів $n_E^{(m)}$ в (3), тобто*

$$P\left[n_E^{(m)}(Z) \neq X \Big| \sum_m\right] - P[n_B(Z) \neq X] = o_P(1) \tag{6}$$

при $m \to \infty$.

Доведення.Припустимо, що $\mathrm{Supp}_+(h) = \{z \in \mathrm{R}^r : h(z) > 0\}$, а також визначимо $S(h_l)$ для множини точок неперервності h_l, $l = 0{,}1$, де

$$M = S\mathrm{upp}_+(h) \bigcap S(h_0) \bigcap S(h_1). \tag{7}$$

Зазначимо, що Z має функцію щільності $z \mapsto h(z) = p_0 h_0(z) + p_1 h_1(z)$, що випливає з теореми Байєса.

Отже, при $P[Z \in M] = 1$ ми маємо

$$P[Z \in \mathrm{R}^r \setminus M] \leqslant$$
$$\leqslant P[Z \in \mathrm{R}^r \setminus \mathrm{Supp}_+(h)] + \sum_{l \in \{0,1\}} P[Z \in \mathrm{R}^r \setminus S(h_l)] \int\limits_{\mathrm{R}^r \setminus \mathrm{Supp}_+(h)} h(z)dz = 0, \tag{8}$$

оскільки $\mathrm{R}^r \setminus S(h_l)$ має нульову міру Лебега, де $l = 0{,}1$ [6].

Далі припустимо, що $z \in M$, де $z \mapsto \theta(z) = p_1 h_1(z)/(p_0 h_0(z) + p_1 h_1(z))$ є неперервною функцією на M. Оцінку $\overline{\theta}_E^{(m)}$ з (3) можна записати у такій формі:

$$\overline{\theta}_E^{(m)} = \sum_{i=1}^{m} X_i D_i^{\gamma(m)}(z) = \frac{1}{N_z^{\gamma(m)}} \sum_{i=1}^{N_z^{\gamma(m)}} X_{z,(i)}, \tag{9}$$

припускаючи, що $X_{z,(i)} = X_{l(z)}$, де $l(z)$ є таким, що $Z_{z,(i)} = Z_{l(z)}$.

Отже,

$$Q^{(m)}(z) := \Omega[\overline{\theta}_E^{(m)}(z) - \theta(z))^2] \leqslant 2Q_1^{(m)}(z) + 2Q_2^{(m)}(z), \tag{10}$$

де

$$Q_1^{(m)}(z) = \Omega\left[\left|\frac{1}{N_z^{\gamma(m)}} \sum_{i=1}^{N_z^{\gamma(m)}} (X_{z,(i)} - \theta(Z_{z,(i)}))\right|^2\right] \tag{11}$$

та

$$Q_2^{(m)}(z) = \Omega\left[\left|\frac{1}{N_z^{\gamma(m)}} \sum_{i=1}^{N_z^{\gamma(m)}} (\theta(Z_{z,(i)}) - \theta(z))\right|^2\right]. \tag{12}$$

Зауважимо, що величину γ_m замінено на γ для спрощення запису. Крім того, $X_{z,(i)} - \theta(Z_{z,(i)})$, $i = 1, \ldots, m$ є нульовими середніми значеннями взаємно незалежних випадкових величин, на що вказує $\sum_{Z}^{(m)}$ для сигма-алгебри, породженої Z_i, $i = 1, \ldots, m$ [7].

В результаті, використовуючи той факт, що $N_z^{\gamma(m)} \geqslant k_m$ майже напевно, ми отримуємо таку рівність:

$$Q_1^{(m)}(z) = \Omega\left[\frac{1}{(N_z^{\gamma(m)})^2} \sum_{i,l=1}^{N_z^{\gamma(m)}} \Omega\left[(X_{z,(i)} - \theta(Z_{z,(i)}))(X_{z,(l)} - \theta(Z_{z,(l)}))| \sum_{Z}^{(m)}\right]\right] =$$
$$= \Omega\left[\frac{1}{(N_z^{\gamma(m)})^2} \sum_{i=1}^{N_z^{\gamma(m)}} \Omega\left[(X_{z,(i)} - \theta(Z_{z,(i)}))^2 | \sum_{Z}^{(m)}\right]\right] \leqslant \Omega\left[\frac{4}{N_z^{\gamma(m)}}\right] \leqslant \frac{4}{k_m} = \mathrm{o}(1) \tag{13}$$

при $m \to \infty$.

Далі, використовуючи нерівність Коші–Шварца, маємо нерівність

$$
\begin{aligned}
&Q_2^{(m)}(z) \leqslant \Omega\left[\frac{1}{N_z^{\gamma(m)}}\sum_{i=1}^{N_z^{\gamma(m)}}(\theta(Z_{z,(i)})-\theta(z))^2\right] = \\
&\quad = \Omega\left[\frac{1}{N_z^{\gamma(m)}}\sum_{i=1}^{N_z^{\gamma(m)}}(\theta(Z_{z,(i)})-\theta(z))^2\Lambda[\|Z_{z,(i)}-z\|\leqslant\tau]\right] + \\
&\qquad + \Omega\left[\frac{1}{N_z^{\gamma(m)}}\sum_{i=1}^{N_z^{\gamma(m)}}(\theta(Z_{z,(i)})-\theta(z))^2\Lambda[\|Z_{z,(i)}-z\|>\tau]\right] \leqslant \\
&\quad \leqslant \sup_{z'\in V_z(\tau)}|\theta(z')-\theta(z)|^2 + 4\Omega\left[\frac{1}{N_z^{\gamma(m)}}\sum_{i=1}^{N_z^{\gamma(m)}}\Lambda[\|Z_{z,(i)}-z\|>\tau]\right] := \\
&\quad := \overline{Q}_2(z;\tau) + \overline{Q}_2^{(m)}(z;\tau)
\end{aligned}
\tag{14}
$$

для $\forall\tau > 0$.

Тому для $\forall\mu > 0$ можна вибрати таке $\tau = \tau(\mu) > 0$, що $\overline{Q}_2(z;\tau(\mu)) < \mu$. Даний вивід слідує з неперервності θ в z. Крім того очевидно, що $Q_2^{(m)}(z;\tau(\mu)) = 0$ для великих m, а оскільки $Q_2^{(m)}(z)$ є o(1), те саме має місце і для $Q^{(m)}(z)$.

Отже, при $m \to \infty$,

$$
\begin{aligned}
&\Omega\left[\left|P\left[\overline{n}_E^{(m)}(Z)\neq X\middle|\sum_m\right]-J_*\right|\right] = \Omega\left[P\left[\overline{n}_E^{(m)}(Z)\neq X\middle|\sum_m\right]-J_*\right], \\
&P[\overline{n}_E^{(m)}(Z)\neq X]-J_* \leqslant 2\Big(\Omega\Big[\overline{\theta}_E^{(m)}(Z)-\theta(Z))^2\Big]\Big)^{1/2} = \mathrm{o}(1),
\end{aligned}
\tag{15}
$$

оскільки $P\left[\overline{n}_E^{(m)}(Z)\neq X\middle|\sum_m\right] \geqslant J_*$ майже напевно. Теорему доведено.

Цитована література

1. *Oja H., Paindaveine D.* Optimal signed-rank tests based on hyperplanes // J. Statist. Plann. Inference. – 2005. – **135**. – P. 307–321.
2. *Chacon J. I., Duong T., Wand M. P.* Asymptotics for general multivariate kernel density derivative estimators // Statist. – 2011. – **21**. – P. 810–837.
3. *Галкін О. А.* Вплив варіацій смуги пропускання на поведінку показника помилкової класифікації ядерного класифікатора // Вісн. Київ. нац. ун-ту ім. Тараса Шевченка. Серія фіз.-мат. науки. – 2014. – № 4. – С. 125–130.
4. *Holmes C. C., Adams N. M.* A probabilistic nearest neighbor method for statistical pattern recognition // J. the Royal Statist. Society. – 2002. – **64**. – P. 298–303.
5. *Jornsten R., Vardi Y., Zhang C. H.* A robust clustering method and visualization tool based on data depth // Statist. data Analysis. – 2002. – P. 354–365.
6. *Rousseeum P. J., Struyf A.* Characterizing angular symmetry and regression symmetry // Statist. Plann. Inference. – 2004. – **122**. – P. 163–170.
7. *Lange T., Mosler K., Mozharovskyi P.* Fast nonparametric classification based on data depth // Statist. Papers. – 2014. – **55**. – P. 53–67.

References

1. *Oja H., Paindaveine D.* J. Statist. Plann. Inference, 2005, **135**: 307–321.
2. *Chacon J. I., Duong T., Wand M. P.* Statist., 2011, **21**: 810–837.
3. *Galkin O. A.* Bulletin of Taras Shevchenko National Univ. of Kyiv, Series Phys. & Math., 2014, No 4: 125–130 (in Ukrainian).
4. *Holmes C. C., Adams N. M.* Journal of the Royal Statistical Society, 2002, **64**: 298–303.
5. *Jornsten R., Vardi Y., Zhang C. H.* Statistical data Analysis, 2002: 354–365.
6. *Rousseeum P. J., Struyf A.* Statist. Plann. Inference, 2004, **122**: 163–170.
7. *Lange T., Mosler K., Mozharovskyi P.* Statist. Papers., 2014, **55**: 53–67.

Надійшло до редакції 17.08.2015

А. А. Галкин

Киевский национальный университет им. Тараса Шевченко

E-mail: galkin.a.o.@gmail.com

Аффинно-инвариантные глубинные классификаторы на основе метода *k*-ближайших соседей

Исследуются глубинные классификаторы на основе метода k-ближайших соседей, которые имеют непараметрическую согласованность при любых непрерывных распределениях. Предложен метод симметризации функции глубины, что обеспечивает центрально-внешнее упорядочение для определения ближайших соседей. Построение симметризации асимптотически гарантирует уникальность наиболее глубокой точки, что решает проблему выпуклой области с бесконечным множеством наиболее глубоких точек. Построенный глубинный классификатор на основе глубинных окрестностей является аффинно-инвариантным, а следовательно, нечувствительным к экстремальным значениям.

Ключевые слова: симметризация, глубинный классификатор, непараметрическая согласованность.

O. A. Galkin

Taras Shevchenko National University of Kiev

E-mail: galkin.a.o.@gmail.com

Affine-invariant depth-based classifiers on the basis of the *k*-nearest neighbors method

Depth-based classifiers on the basis of the k-nearest neighbors method are studied with nonparametric consistency for any continuous distribution. The method of symmetrization of a depth function is proposed, providing a centrally external ordering to determine the nearest neighbors. The construction of a symmetrization asymptotically guarantees the uniqueness of the deepest point that solves the problem of a convex domain with an infinite set of the deepest points. The constructed depth-based classifier based on the depth-based neighborhoods is affine invariant and, therefore, insensitive to extreme values.

Keywords: symmetrization, depth-based classifier, nonparametric consistency.