

10. Wilson D. L. Asymptotic properties of nearest neighbor rules using edited data / D. L. Wilson // IEEE Transactions on Systems, Man, Cybernetics. – 1972. – Vol. 2, № 3. – P. 408–421.
11. Wilson D. R. Reduction techniques for instancebased learning algorithms / D. R. Wilson, T. R. Martinez // Machine Learning. – 2000. – Vol. 38, № 3. – P. 257–286.

S. Subbotin

THE AUTOMATIC SAMPLE EXTRACTION FOR NEURAL NETWORK MODEL BUILDING

The sample selection method is proposed, which for the original sample determines the individual instance significance, followed by successive increase of a subsample, selecting the most highly informative individually instances in each class, and excluding instances that are redundant or worse a classification. This allows to automate the sample analysis, to reduce the training data dimensionality, to reduce the time and to provide the acceptable accuracy of neural network training. The experiments to investigate the proposed method are conducted. Their results allow recommend method to use in practice for the diagnosis and pattern recognition.

Keywords: sample, instance selection, data reduction, neural network, data dimensionality reduction.

Матеріал надійшов 01.09.2014

УДК 519.7

Галкін О. А.

НЕПАРАМЕТРИЧНІ ОЦІНКИ УСЕРЕДНЕНИХ ЯДЕРНИХ ВІДОБРАЖЕНЬ УМОВНИХ РОЗПОДІЛІВ ДЛЯ ЗАДАЧ РОЗПІЗНАВАННЯ ОБРАЗІВ

Статтю присвячено непараметричним оцінкам усереднених ядерних відображень умовних розподілів, що є неявними відображеннями розподілу в потенційно нескінченновимірний простір характеристик, а також комплексному ядерному підходу для розв'язання широкого класу задач розпізнавання образів. Ключова ідея полягає у відображенні умовного розподілу в нескінченновимірний простір характеристик з використанням функції ядра. Запропонований підхід може бути використаний для побудови більш простих та ефективних статистик для оцінки такого неперервного мультимодального розподілу, як функція глибини.

Ключові слова: ядерне відображення, функція ядра, оцінка щільності.

Вступ

Розглядаючи неперервні генеральні сукупності, припустимо, що задано випадкову величину Z , генеральну сукупність H , розподіл $D(Z)$, а також

щільність $d(Z)$. Будемо вважати, що випадкова величина X також належить генеральній сукупності H , а також припустимо, що запропонований підхід має місце лише у випадку, коли Z та X належать різним генеральним сукупностям.

Гільбертовий простір функцій $f: H \mapsto \mathbb{R}$ зі скалярним добутком $[\cdot, \cdot]_{\mathfrak{Z}}$ є Гільбертовим простором відтворюваного ядра \mathfrak{Z} на H з ядром $O(z, z')$, елемент $O(z, \cdot)$ якого задовольняє властивість $[f(\cdot), O(z, \cdot)]_{\mathfrak{Z}} = f(z)$. Враховуючи рівність

$$[O(z, \cdot), O(z', \cdot)]_{\mathfrak{Z}} = O(z, z'),$$

як скалярний добуток можна розглядати оцінку функції f в будь-якій точці $z \in H$. Крім того, $O(z, \cdot)$ може виступати в ролі неявного характеристичного відображення $\theta(z)$, де $O(z, z') = [\theta(z), \theta(z')]_{\mathfrak{Z}}$. Зазначимо, що функції ядра в \mathbb{R}^n включають поліноміальне ядро $O(z, z') = ([z, z'] + c)^l$ та гаусівську радіально-базисну функцію ядра $O(z, z') = \exp(-\sigma \|z - z'\|^2)$, тому, поки для них визначені функції ядра, запропонований нижче підхід може бути узагальнений для широкого спектра типів даних.

Виклад основного матеріалу

Метод усередненого ядерного відображення полягає в поелементному розподілі ймовірностей у Гільбертовому просторі відтворюваного ядра, що пов'язаний з функцією ядра

$$\sigma_z = \Theta_z[\theta(Z)] = \int_H \theta(z) dD(z).$$

У цьому випадку розподіл відображається в точку в потенційно нескінченновимірному та неявному просторі характеристик, тобто в очікуване характеристичне відображення. Середнє значення усередненого відображення σ_z має таку властивість, що очікування будь-якої функції f Гільбертового простору відтворюваного ядра може бути оцінене як скалярний добуток в \mathfrak{Z} , де $[\sigma_z, f] = \Theta_z[f(Z)]$, $\forall f \in \mathfrak{Z}$. Зазначимо, що середнє значення усередненого відображення відрізняється від ядерних оцінок щільності, де щільність є згорткою зі згладжувальним ядром $\tilde{O}(z, z')$. Крім того, ця щільність є дійсною ймовірнісною щільністю, отриманою за допомогою ядерних оцінок щільності $Y(z') = \Theta_z[\tilde{O}(Z, z')]$, де згладжувальне ядро не може бути додатновизначеним, а $\tilde{O}(z, z') \neq [\tilde{\theta}(z), \tilde{\theta}(z')]_{\mathfrak{Z}}$. Коли функція згладжувального ядра є додатновизначеною (наприклад, гаусівська радіально-базисна функція ядра $\tilde{O}(z, z') = \exp(-\sigma \|z - z'\|^2)$), пропускна здатність ядра в ядерних оцінках щільності (а отже, і простір характеристик) часто змінюється залежно від кількості спостережуваних точок, а тому не може бути оцінена як усереднене

відображення розподілів у фіксований простір характеристик [9, р. 1–22].

Усереднене ядерне відображення може бути узагальнене до спільних розподілів двох або більше змінних шляхом використання простору характеристик тензорного добутку. Як наслідок, можна відобразити спільний розподіл двох змінних Z та X у простір характеристик тензорного добутку $\mathfrak{Z} \times \mathfrak{Z}$ таким чином:

$$\Delta_{ZX} = \Theta_{ZX}[\theta(Z) \times \theta(X)] = \int_{H \times H} \theta(z) \times \theta(x) dD(z, x),$$

де припускається, що дві змінні ділять між собою одну й ту саму генеральну сукупність H та ядро O , а характеристики тензорного добутку задовольняють таку умову:

$$[\theta(z) \times \theta(x), \theta(z') \times \theta(x')]_{\mathfrak{Z} \times \mathfrak{Z}} = O(z, z') O(x, x').$$

Як окремі випадки усередненими ядерними відображеннями є граничний ймовірнісний вектор дискретної змінної Z та таблиця ймовірностей спільного розподілу дискретних змінних Z та X . Використовуючи дельта-ядро Кронекера $O(z, z') = \mu(z, z')$, припустимо, що $z, x \in \{1, \dots, M\}$. Тоді відповідне характеристичне відображення $\theta(z)$ є стандартним базисом β_z в \mathbb{R}^M . Отже, ми маємо, що

$$\begin{pmatrix} D(z=1) \\ \dots \\ D(z=M) \end{pmatrix} = \Theta_z[\beta_z] = \sigma_z,$$

та

$$\begin{pmatrix} D(z=s, x=t) \end{pmatrix} = \Theta_{ZX}[\beta_z \times \beta_x] = \Delta_{ZX}.$$

За стандартною еквівалентністю між тензором та лінійним відображенням спільні усереднені ядерні відображення можна розглядати як нецентрований крос-коваріаційний оператор $\Delta_{ZX}: \mathfrak{Z} \mapsto \mathfrak{Z}$ [8, р. 394–416]. Коваріація функцій $f, g \in \mathfrak{Z}$ може бути обчислена як $\Theta_{ZX}[f(Z)g(X)] = [f, \Delta_{ZX}g]_{\mathfrak{Z}}$. Оскільки Δ_{ZX} розглядався як оператор, а тепер як елемент у просторі тензорного добутку, ця коваріація може бути обчислена як $[f \times g, \Delta_{ZX}]_{\mathfrak{Z} \times \mathfrak{Z}}$. Таким же чином можуть бути обчислені вирази $\Delta_{ZZ} = \Theta_z[\theta(Z) \times \theta(Z)]$ та $\Delta_{(ZZ)X} = \Theta_x[\theta(Z) \times \theta(Z) \times \theta(X)]$. Останній вираз можна розглядати як лінійний оператор з \mathfrak{Z} в $\mathfrak{Z} \times \mathfrak{Z}$.

Усереднене ядерне відображення є ін'єкційним для характеристичних ядер $[\cdot, \cdot]$, тобто вони будуть відображатися у дві різні точки в просторі характеристик за умови, якщо розподіли

$P(Z)$ та $W(Y)$ є різними. Більшість часто використовуваних ядер є характерними, наприклад, гаусівська радіально-базисна функція ядра $\exp(-\sigma\|z-z'\|^2)$ та ядро Лапласа $\exp(-\sigma\|z-z'\|)$. Це означає, що, якщо розподіл відображено за допомогою цих ядер, відстань відображень у просторі характеристик буде індикатором ідентичності розподілів [1, р. 185–240].

Незважаючи на те, що отримання істинного вихідного розподілу $D(Z)$ є практично неможливим, його усереднені ядерні відображення можна оцінити за допомогою кінцевого вибіркового середнього значення. Маючи вибірку $V_Z = \{z_1, \dots, z_n\}$ розміру n , елементи якої є незалежними та однаково розподіленими випадковими величинами з $D(Z)$, емпіричне усереднене ядерне відображення матиме такий вигляд:

$$\tilde{\sigma}_Z = \frac{1}{n} \sum_{i=1}^n \theta(z_i).$$

Усереднені ядерні відображення спільних розподілів успадковують попередні дві властивості загальних відображень: ін'єктивність та тривіальну емпіричну оцінку. За наявності n пар навчальних прикладів $V_{ZX} = \{(z_1, x_1), \dots, (z_n, x_n)\}$, що є незалежними та однаково розподіленими випадковими величинами з $D(Z, X)$, коваріаційний оператор Δ_{ZX} може бути оцінений як

$$\Delta_{ZX} = \frac{1}{n} \sum_{i=1}^n \theta(z_i) \times \theta(x_i).$$

Завдяки ядерному прийому більшість обчислень, необхідних для статистичного виходу, що використовує усереднені ядерні відображення, можуть бути зменшені до маніпуляцій з матрицею Грама. Входи матриці Грама K відповідають ядерному значенню між точками z_i та z_j , тобто $K_{ij} = O(z_i, z_j)$, а тому її розмір визначається кількістю точок у вибірці (так само матриця Грама G має входи $G_{ij} = O(x_i, x_j)$). Розмірність матриці Грама є набагато меншою, ніж розмірність простору характеристик (що може бути нескінченним). Це дозволяє використовувати ефективні непараметричні методи з використанням ядерних відображень. Наприклад, емпіричний багатомодульний розподіл може бути обчислений шляхом використання ядерних оцінок, а саме:

$$\begin{aligned} \Sigma(D, W) &= \left\| \frac{1}{n} \sum_{i=1}^n \theta(z_i) - \frac{1}{n} \sum_{i=1}^n \theta(x_i) \right\|_{\mathfrak{S}}^2 = \\ &= \frac{1}{n^2} \sum_{i,j=1}^n (O(z_i, z_j) + O(x_i, x_j) - 2O(z_i, x_j)). \end{aligned}$$

Для порівняння: відстань L_2 між ядерними оцінками щільності є такою:

$$\begin{aligned} &\int_{\mathfrak{S}} (\nu(z) - \nu'(z))^2 dz = \\ &= \frac{1}{n^2} \int_{\mathfrak{S}} \sum_{i,j=1}^n \left(\begin{aligned} &\tilde{O}(z_i, z) \tilde{O}(z_j, z) + \tilde{O}(x_i, z) \tilde{O}(x_j, z) - \\ &- 2\tilde{O}(z_i, z) \tilde{O}(x_j, z) \end{aligned} \right) dz, \end{aligned}$$

де $\nu(z) = \frac{1}{n} \sum_{i=1}^n \tilde{O}(z_i, z)$, а $\nu'(z) = \frac{1}{n} \sum_{i=1}^n \tilde{O}(x_i, z)$.

Більше того, можна продемонструвати, що дво-вибірковий тест на основі відстані L_2 між ядерними оцінками щільності має меншу ефективність проти локального зсуву від нульової гіпотези, ніж багатомодульний розподіл, через стиснуту ядерну пропускну здатність із збільшенням обсягу вибірки. Нарешті, такий гіперпараметр ядерних функцій, як смуга пропускання σ в гаусівській радіально-базисній функції ядра $\exp(-\sigma\|z-z'\|^2)$, може бути обраний для максимізації тестової ефективності та мінімізації ймовірності типу помилка II у дво-вибіркових тестах (тобто ймовірність помилково визначених D та W є такою ж, як і у випадку, коли вони є фактично різними) [2]. Якщо обсяг вибірки є достатньо великим, обчислення в методах усередненого ядерного відображення може бути достатньо громіздким. У цьому випадку ефективним рішенням є використання такої низькорангової апроксимації матриці Грама, як неповна факторизація Холецького, що є досить ефективним методом у зменшенні обчислювальних витрат ядерних методів без втрати точності апроксимації.

Незважаючи на те, що усереднені ядерні відображення розподілів забезпечують потужну структуру для роботи з низкою складних багатомірних непараметричних задач, залишається ще багато питань, які потрібно вирішити щодо використання цих відображень для виведення. Далі представлено ядерний непараметричний метод для умовних розподілів, що дає змогу враховувати умовні відношення виведення, а також забезпечити уніфіковану ядерну структуру для правила суми, правила добутку та байєсівського правила.

Усереднене ядерне відображення умовного розподілу $D(X|Z)$ визначається як

$$\sigma_{x|z} = \Theta_{x|z}[\theta(X)] = \int_{\mathfrak{S}} \theta(x) dD(x|z).$$

На основі цього відображення умовне математичне очікування функції $g \in \mathfrak{S}$ може бути обчислене як $\Theta_{x|z}[g(X)] = [g, \sigma_{x|z}]_{\mathfrak{S}}$. Це можна порівняти з властивістю відображення

середнього значення, де безумовне математичне очікування функції можна записати у вигляді скалярного добутку з відображенням [6]. Усереднене відображення умовного розподілу не є єдиним елементом у Гільбертовому просторі відтворюваного ядра, але замість цього враховується сімейство точок у Гільбертовому просторі відтворюваного ядра, кожна з яких індексується фіксованим значенням x умовної змінної Z . Зауважимо, що це є лише фіксацією Z до певного значення z , де ми будемо в змозі отримати один елемент Гільбертового простору відтворюваного ядра $\sigma_{x|z} \in \mathfrak{S}$. Інакше кажучи, потрібно визначити оператор, що позначається як $\Delta_{x|z}$, який може приймати як вхідні дані z , а на виході отримувати відповідне відображення. Тобто необхідно, щоб задовольнялась така умова:

$$\sigma_{x|z} = \Delta_{x|z} \theta(z).$$

На основі співвідношення між умовним математичним сподіванням та коваріаційним оператором, а також за припущення, що $\Theta_{x|z}[g(X)] \in \mathfrak{S}$,

$$\Delta_{x|z} = \Delta_{xz} \Delta_{zz}^{-1},$$

а тому $\sigma_{x|z} = \Delta_{xz} \Delta_{zz}^{-1} \theta(z)$ задовольняє умову щодо $\sigma_{x|z} = \Delta_{x|z} \theta(z)$. Визначення умовного оператора усередненого відображення в $\Delta_{x|z} = \Delta_{xz} \Delta_{zz}^{-1}$ є досить загальним, а умовна ймовірність $D(X|Z)$ для дискретних змінних є окремим випадком [1, р. 1–36]. Використовуючи дельта-ядро Кронекера та конструкцію, аналогічну

$$\begin{pmatrix} D(z=1) \\ \dots \\ D(z=M) \end{pmatrix} = \Theta_z[\beta_z] = \sigma_z,$$

та

$$\begin{pmatrix} D(z=s, x=t) \end{pmatrix} = \Theta_{zx}[\beta_z \times \beta_x] = \Delta_{zx},$$

ми можемо отримати таблицю умовних імовірностей за таким співвідношенням:

$$\underbrace{\begin{pmatrix} D(x=s | z=t) \end{pmatrix}}_{\Delta_{x|z}} = \underbrace{\begin{pmatrix} D(z=1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & D(z=M) \end{pmatrix}}_{\Delta_{zz}^{-1}}^{-1} \underbrace{\begin{pmatrix} D(x=s, z=t) \end{pmatrix}}_{\Delta_{xz}}$$

Враховуючи множину даних $V_{zx} = \{(z_1, x_1), \dots, (z_n, x_n)\}$ розміру n , елементи якої є незалежними та однаково розподіленими випадковими величинами з $D(Z, X)$, ми будемо оцінювати умовний ядерний оператор відображення як

$$\Delta_{x|z} = \Phi(K + \tau I)^{-1} \Upsilon^T,$$

де $\Phi = (\theta(x_1), \dots, \theta(x_n))$ та $\Upsilon = (\theta(x_1), \dots, \theta(x_n))$ є неявно сформованими матрицями характеристик, а $K = \Upsilon^T \Upsilon$ є матрицею Грама для вибірок зі змінної Z . Крім того, необхідно мати додатковий параметр регуляризації τ , щоб уникнути явища перенавчання. У результаті, $\sigma_{x|z} = \Delta_{x|z} \theta(z)$ стає зваженою сумою характеристично відображених точок з X , а усереднене ядерне відображення умовного розподілу знаходиться таким чином:

$$\sigma_{x|z} = \sum_{i=1}^n \omega_i(z) \theta(x_i) = \Phi \omega(z),$$

де $\omega(x) = (\omega_1(x), \dots, \omega_n(x))^T = (K + \tau I)^{-1} K_{:,z}$, а $K_{:,z} = (O(z, Z_1), \dots, O(z, Z_n))^T$.

Емпірична оцінка усередненого умовного відображення є аналогічною оцінці звичайного усередненого відображення, що задається рівнянням $\sigma_z = \Theta_z[\theta(Z)] = \int_H \theta(z) dD(z)$. Різниця полягає в тому, що замість рівномірних ваг, $1/n$, застосовуються нерівномірні ваги $\omega_i(x)$ на спостереження, які, своєю чергою, визначаються значенням x змінної кондиціонування [4]. Ці нерівномірні ваги відображають ефект кондиціонування на усереднених відображеннях. Крім того, ця емпірична оцінка сходиться до своєї генеральної сукупності за нормою $\|\sigma_{x|z} - \sigma_{x|z}\|_{\mathfrak{S}}$ з коефіцієнтом $O_p(n^{-\frac{1}{4}})$ при зменшенні регуляризації τ з коефіцієнтом $O(n^{-\frac{1}{2}})$ [5]. За наявності відповідних припущень на спільному розподілі Z і X можуть бути отримані кращі коефіцієнти. Відповідно, оцінка умовного оператора відображення в

$$\sigma_{x|z} = \sum_{i=1}^n \omega_i(z) \theta(x_i) = \Phi \omega(x)$$

буде досить суттєво відрізнятися від умовної ядерної оцінки щільності

$$\Lambda(x|z) = \frac{\sum_{i=1}^n O(x_i, x) O(z_i, z)}{\sum_{i=1}^n O(z_i, z)},$$

що має певні недоліки для багатовимірних даних, та в областях, де значення $d(z)$ є невеликим. На відміну від цього, умовні оператори усередненого відображення не оцінюють безпосередньо щільність, а тому уникають цих проблем. При заданій функції $g(x) = \sum_{i=1}^n \rho_i O(\tilde{x}_i, x)$ з Гільбертового

простору відтвореного ядра ми можемо оцінити її очікуване значення по відношенню до умовного розподілу, використовуючи $\sigma_{X|z}$, за допомогою матричних операцій, а саме:

$$\Theta_{X|z}[g(X)] \approx [g, \sigma_{X|z}]_{\mathfrak{S}} = \sum_{i=1}^{\tilde{n}} \sum_{j=1}^n \rho_i \omega_j(z) O(\tilde{x}_i, x_j) = \rho^T \tilde{G} (K + \tau I)^{-1} K_{:z},$$

де \tilde{G} є матрицею Грама з елементами $\tilde{G}_{ij} = O(\tilde{x}_i, x_j)$, а $\rho = (\rho_1, \dots, \rho_{\tilde{n}})^T$. На відміну від цього, умовна ядерна оцінка щільності вимагає інтегрування для оцінки за H , тобто:

$$\Theta_{X|z}[g(X)] \approx \int_H g(x) \Lambda(x|z) dx = \int_H \frac{\sum_{i=1}^{\tilde{n}} \sum_{j=1}^n \rho_i O(\tilde{x}_i, x) O(x_j, x) O(z_j, z)}{\sum_{j=1}^n O(z_j, z)} dx,$$

що може бути досить складним для великої розмірності X . Умовний оператор відображення альтернативно може бути знайдений як розв'язок функції регресійної задачі найменших квадратів [3]:

$$\Delta_{X|Z} = \arg \min_{\Delta: \mathfrak{S} \rightarrow \mathfrak{S}} \sum_{i=1}^n \|\theta(x_i) - \Delta \theta(z_i)\|_{\mathfrak{S}}^2 + \tau \|\Delta\|_{HSH}^2,$$

де $\|\Delta\|_{HSH}^2$ позначає норму Гільберта – Шмідта (узагальнена норма Фробеніуса) оператора Δ . Одним з практичних наслідків цього зв'язку є те, що гіперпараметри в умовних усереднених відображеннях, як-от параметри для ядер за змінною Z та параметр регуляризації τ , можуть бути обрані за допомогою методу перехресної перевірки на основі регресії.

Список літератури

- Berlinet A. Reproducing Kernel Hilbert Spaces in Probability and Statistics / A. Berlinet, C. Thomas-Agnan. – NY : Kluwer, 2004. – 353 p.
- Fine S. Efficient SVM training using low-rank kernel representation / S. Fine, K. Scheinberg // Journal of Machine Learning Research. – 2001. – Vol. 2. – P. 243–264.
- Girolami M. Probability density estimation from optimally condensed data samples / M. Girolami, C. He // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2003. – Vol. 25 (10). – P. 1253–1264.
- Hall P. Cross-Validation and the Estimation of Conditional Probability Densities / P. Hall, J. Racine, Q. Li // Journal of the American Statistical Association. – 2004. – Vol. 99. – P. 1015–1026.
- Hansen B. E. Uniform Convergence Rates for Kernel Estimation with Dependent Data / B. E. Hansen. – Econometrics Theory. – 2008. – Vol. 24. – P. 726–748.
- Ihler A. Particle belief propagation / A. Ihler, D. McAllester // Proc. Intl. Conference on Artificial Intelligence and Statistics. – 2009. – P. 256–263.
- Raykar V. C. Very fast optimal bandwidth selection for univariate kernel density estimation : Technical Report CS-TR-4774 / V. C. Raykar, R. Duraiswami. – University of Maryland, CollegePark, 2005.
- Schölkopf B. Kernel Methods in Computational Biology / B. Schölkopf, K. Tsuda, J.-P. Vert. – Cambridge, MA : MIT Press, 2004.
- Silverman B. W. Density estimation for Statistical and Data Analysis : Monographs on statistics and applied probability / B. W. Silverman. – London : Chapman and Hall, 1986.

O. Galkin

NON-PARAMETRIC ESTIMATES OF THE MEAN KERNEL MAPPINGS OF CONDITIONAL DISTRIBUTIONS FOR PATTERN RECOGNITION PROBLEMS

This article focuses on nonparametric estimates of the mean kernel mappings of conditional distributions that are implicit mappings of distribution in a potentially infinite dimensional space of features, as well as complex kernel approach for solving a wide class of pattern recognition problems. The key idea is the mapping of conditional distribution in infinite space of features using the kernel function. The proposed approach can be used to construct a simple and effective evaluation of the statistics for continuous multimodal distribution as a depth function.

Keywords: kernel mapping, kernel function, kernel density estimate.

Матеріал надійшов 03.03.2015