仮想的な 0 近傍への外挿とその収束レートについて *

奥野彰文 [1,3], 下平英寿 [2,3]

[1] 統計数理研究所, [2] 京都大学大学院 情報学研究科, [3] 理化学研究所 AIP センター

1 はじめに

分類問題における教師あり学習に幅広く用いられる手法の一つに k 近傍法 (k-Nearest Neighbour, 略称 k-NN; Fix and Hodges (1951)) がある．k 近傍法ではまずクエリ近傍の k 個のデータベクトルを検索し，対応するラベルの平均によりクエリの各ラベル確率を予測する．最大の確率を持つラベルを出力するプラグイン型の分類を行うと，k 近傍法の統計的一致性，すなわち誤分類確率が最適値に収束することが Cover and Hart (1967) などにより示されており，またその収束の速さ (収束レート) が Chaudhuri and Dasgupta (2014) などにより示されている．

一方で，任意のラベル確率予測器を用いてプラグイン型の分類を行う場合の最適な収束レートが Audibert and Tsybakov (2007) により与えられており，ラベルの条件付き期待値が非常に滑らかな場合には k 近傍法は最適レートを達成できない．

k 近傍法にはバイアスとバリアンスのトレードオフがある：k を大きくするとクエリから遠いデータベクトルのラベルが考慮されるようになり，バイアスが増大してしまう．本研究では，いくつかの k について k 近傍法を行い，予測されたラベル確率を $k = 0$ に仮想的に外挿することで漸近的なバイアスを減少させ，最適な収束レートを達成する**マルチスケール k 近傍法**を提案する．

2 問題設定

$\mathcal{X} \subset \mathbb{R}^d$ を非空なコンパクト集合とし，$(X, Y) \in \mathcal{X} \times \{0, 1\}$ を分布 $\mathbb{Q}$ から生成される確率変数とする．サンプル $\mathcal{D}_n := \{(X_i, Y_i)\}_{i=1}^n$ とクエリ (X_*, Y_*) は独立に分布 $\mathbb{Q}$ に従うとする．$\mathcal{D}_n, X_*$ を用いてクエリとなるデータベクトル $X_* \in \mathcal{X}$ のラベル $Y_* \in \{0, 1\}$ を予測する分類器 $\hat{g}_n : \mathcal{X} \to \{0, 1\}$ を学習し評価する．

分類器の評価には誤分類確率 $L(g) := \mathbb{P}_{X_*, Y_*}(g(X_*) \neq Y_*)$ を用いた excess risk

$$\mathcal{E}(\hat{g}_n) := \mathbb{E}_{\mathcal{D}_n}(L(\hat{g}_n)) - \inf_{g:\mathcal{X}\to\{0,1\}} L(g)$$

を用いる．Excess risk $\mathcal{E}(\hat{g}_n)$ のサンプル数 n についてのオーダーを**収束レート**と呼ぶ．

3 k 近傍法

クエリからのユークリッド距離により添え字を並べ替える：$\|X_{(1)} - X_*\|_2 \leq \|X_{(2)} - X_*\|_2 \leq \cdots \leq \|X_{(n)} - X_*\|_2$．ユーザの指定するパラメータ $k \in \mathbb{N}$ と和が 1 となる重み $w_1, w_2, \ldots, w_k \geq 0$ を用いた

$$\hat{\eta}_{k,\boldsymbol{w}}^{(k\mathrm{NN})}(X_*) = \sum_{i=1}^{k} w_i Y_{(i)}$$

を重み付き k 近傍推定量 (k-NN estimator) と呼び，$w_1 = w_2 = \cdots = w_k = k^{-1}$ の場合単純に k 近傍推定量と呼ぶ．プラグイン型の分類器 $\hat{g}_{k,\boldsymbol{w}}^{(k\mathrm{NN})}(X_*) := 1\left(\hat{\eta}_{k,\boldsymbol{w}}^{(k\mathrm{NN})}(X_*) \geq 1/2\right)$ を特に重み付き k 近傍分類器 (weighted k-NN classifier) と呼ぶ．以降ではクエリ X_* を固定し記号を省略する．

4 マルチスケール k 近傍法

図 1には各 k での k 近傍推定量を示した．k が大きいほど k 近傍推定量の分散は小さいが，青線で示した真値との乖離 (バイアス) が大きくなる．

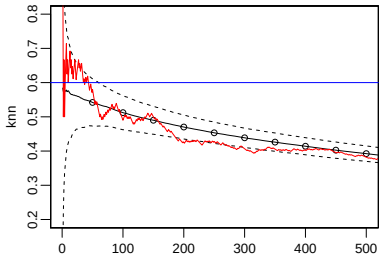

図 1: 赤線：k 近傍推定量，青線：真値 $\eta(X_*) = \mathbb{E}(Y \mid X_*)$，黒線と破線はモンテカルロシミュレーションにおける k 近傍推定量の平均と標本標準偏差を表す．

本研究では，$1 \leq k_1 < k_2 < \cdots < k_V \leq n$ $(V \in \mathbb{N})$ を用いて k 近傍推定量 $\hat{\eta}_{k_1}^{(k\mathrm{NN})}, \hat{\eta}_{k_2}^{(k\mathrm{NN})}, \ldots, \hat{\eta}_{k_V}^{(k\mathrm{NN})}$ を計算し，$(k =)0$ 近傍へ外挿することでバイアスを減少させる**マルチスケール k 近傍法**を提案する．より具体的な手続きとしては

(1) 半径 $r_k := \|X_{(k)} - X_*\|_2$ $(k = k_1, \ldots, k_V)$ を計算．
(2) 半径 r_{k_v} を介した回帰関数 $f(r_{k_v}; \theta)$ を用いて k 近傍推定量 $\hat{\eta}_{k_v}^{(k\mathrm{NN})}$ を予測する：

$$\hat{\theta}_{\boldsymbol{k}} := \arg\min_{\theta\in\Theta} \sum_{v=1}^{V} \{\hat{\eta}_{k_v}^{(k\mathrm{NN})} - f(r_{k_v}; \theta)\}^2.$$

(3) $r_0 = 0$ として，$k = 0$ に外挿する：

$$\hat{\eta}_{\boldsymbol{k}}^{(\mathrm{MS}k\mathrm{NN})} := f(0; \hat{\theta}_{\boldsymbol{k}}).$$

対応するプラグイン型のマルチスケール k 近傍分類器を $\hat{g}_{\boldsymbol{k}}^{(\mathrm{MS}k\mathrm{NN})}(X_*) := 1\left(\hat{\eta}_{\boldsymbol{k}}^{(\mathrm{MS}k\mathrm{NN})}(X_*) \geq 1/2\right)$ とする．

* 本稿は Okuno and Shimodaira (2020) を基にしている．

次節では次数が偶数の項のみで構成された多項式

$$f(r;\theta) = \theta_0 + \theta_1 r^2 + \theta_2 r^4 + \cdots + \theta_C r^{2C} \quad (1)$$

を回帰関数として用いると，近傍法の収束レートが改善することを紹介する．

5 理論解析

本節では，マルチスケール k 近傍法が通常の k 近傍法の収束レートを改善し，最適レートを達成することを示す．最初に，条件付き期待値 $\eta(x) = \mathbb{E}[Y \mid X = x]$ に関するいくつかの条件を説明する．

定義 1 (α-margin 条件)．ある定数 $L_\alpha \geq 0, \tilde{t} > 0, \alpha > 0$ が存在して $\mathbb{P}(|\eta(X) - 1/2| \leq t) \leq L_\alpha t^\alpha$ $(\forall t \in (0, \tilde{t}], X \in \mathcal{X})$ とできるとき η は α-margin 条件を満たすという．

定義 2 (β-Hölder 条件)．$\mathcal{T}_{q,X_*}[\eta]$ を点 $X_* \in \mathcal{X}$ での関数 η の $q\,(\in \mathbb{N}_0)$ 次テイラー展開とする．ある定数 $L_\beta > 0, \beta > 0$ が存在して $|\eta(X) - \mathcal{T}_{\lfloor\beta\rfloor,X_*}[\eta](X)| \leq L_\beta \|X - X_*\|^\beta$ とできるとき $\eta \in C^{\lfloor\beta\rfloor}(\mathcal{X})$ は β-Hölder 条件を満たすという．

定義 3 (γ-neighbour average smoothness 条件)．集合 $B \subset \mathbb{R}^d$ について定義される関数 $\eta^{(\infty)}(B) := \mathbb{E}(Y \mid X \in B)$ と中心 $X \in \mathcal{X}$ で半径 r の球 $B(X;r)$ について，ある定数 $L_\gamma > 0, \gamma > 0$ が存在して $|\eta^{(\infty)}(B(X;r)) - \eta(X)| \leq L_\gamma r^\gamma$ を満たすとき η は γ-neighbour average smooth であるという．

一般に α, β, γ が大きいほど分類器の収束が早くなる．最後に，X の密度関数について以下の条件を定義すると定理 5が成り立つ．

定義 4 (Strong density assumption)．X の密度関数 μ について，ある定数 $\mu_{\min}, \mu_{\max} \in (0, \infty)$ が存在して $\mu_{\min} \leq \mu(X) \leq \mu_{\max}$ $(\forall X \in \mathcal{X})$ とできるとき μ は strong density assumption (SDA) を満たすという．

定理 5 (k 近傍分類器の収束レート; Chaudhuri and Dasgupta (2014) 定理 4(b))．集合 $\mathcal{X}$ がコンパクトであり，η が α-margin と γ-neighbour average smoothness 条件を満たすとする．SDA が成り立つとし，$k_* := k_n \asymp n^{2\gamma/(2\gamma+1)}$ とすると $\mathcal{E}(\hat{g}_k^{(k\mathrm{NN})}) = O(n^{-(1+\alpha)\gamma/(2\gamma+d)})$.

定理 5は γ-neighbour average smoothness 条件を仮定するが，代わりに β-Hölder 条件を仮定すると，いくつかの条件の下で $\gamma = \min\{\beta, 2\}$ が示せる (Okuno and Shimodaira (2020) 定理 1)．つまり $\beta > 2$ として β-Hölder 条件を仮定すると

$$\mathcal{E}(\hat{g}_k^{(k\mathrm{NN})}) = O(n^{-2(1+\alpha)/(4+d)})$$

となり，β がいくら大きくなっても収束レートがバウンドされてしまう．これは局所線形回帰 (Tsybakov, 2009) と同様の現象である (Hall and Kang, 2005)．

次に，β-Hölder 条件の下でマルチスケール k 近傍法の収束レートを以下の定理 6に示す．

定理 6 (マルチスケール k 近傍分類器の収束レート; Okuno and Shimodaira (2020) 定理 2)．$\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_V) \in \mathbb{R}^V$ を $\ell_1 = 1 < \ell_2 < \cdots < \ell_V < \infty$ なるベクトルとし，$k_{1,n} \asymp n^{2\beta/(2\beta+d)}$ かつ $k_{v,n} \asymp \min\{k \in [n] \mid \|X_{(k)} - X_*\|_2 \geq \ell_v \|X_{(k_{1,n})} - X_*\|_2\}$ とする．回帰関数として (1) を用いる．(i) $\mu, \mu\eta$ がそれぞれ β-Hölder 条件を満たし，(ii) SDA を仮定し，かつ (iii) $C := \lfloor\beta/2\rfloor \leq V - 1$ とすると，推定量が発散しないための条件 (Okuno and Shimodaira (2020) (C-3)) の下で

$$\mathcal{E}(\hat{g}_{\boldsymbol{k}}^{(\mathrm{MS}k\mathrm{NN})}) = O(n^{-(1+\alpha)\beta/(2\beta+d)}). \quad (2)$$

式 (2) で示したマルチスケール k 近傍法の収束レートは，任意のプラグイン型の分類器に関する最適レート (Audibert and Tsybakov, 2007) と一致する．

6 まとめと議論

以上より，提案したマルチスケール k 近傍法は k 近傍法の収束レートを改善し，最適レートを達成することを示した．最後に，本節では提案法と局所多項式回帰，重み付き k 近傍法との比較についての議論をまとめる．

6.1 局所多項式回帰との比較

Audibert and Tsybakov (2007) では，局所多項式回帰 (Tsybakov, 2009) を用いたプラグイン型の分類器が最適レートを達成することを示している．局所多項式回帰はクエリ $X_* \in \mathcal{X} \subset (\mathbb{R}^d)$ の周辺での η のテイラー展開を多項式で推定するため，$1 + d + d^2 + \cdots + d^C$ 個の係数を推定する必要があるが，提案したマルチスケール k 近傍法では (1) に現れる $1 + C$ 個の項の係数を推定するだけでよい．

6.2 重み付き k 近傍法との比較

単純な k 近傍法と回帰を組み合わせ容易に実装することができるマルチスケール k 近傍法は，別の解釈として，回帰を介し間接的に重み付き k 近傍法の**実数値**の重み $w_1, \ldots, w_k \in \mathbb{R}$ を推定しているとみなせる．

重み付き k 近傍法では通常，非負の重みのみを考える．非負の重みのみ考える場合，最適な重みを用いても，収束レートが重みを用いない k 近傍法と同じになってしまうことが知られている (Samworth, 2012)．Samworth (2012) ではさらに excess risk のテイラー展開を最適化する実数の重みを用いれば最適レート (2) が達成できることを示している．

提案したマルチスケール k 近傍法は回帰を介して，Samworth (2012) は excess risk のテイラー展開の最適化を介して重みを決定することから，これら 2 つの手法で得られる実数値の重みは同一ではないが，どちらも最適なレートを達成する．しかし Samworth (2012) の方法ではテイラー展開を重みに関して最適化する必要があり，最適な重みの計算が煩雑になってしまう．提案法には回帰を介して容易に同等な計算を可能にする利点がある．

参考文献

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633.

Chaudhuri, K. and Dasgupta, S. (2014). Rates of Convergence for Nearest Neighbor Classification. In *Advances in Neural Information Processing Systems 27*, pages 3437–3445. Curran Associates, Inc.

Cover, T. and Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Fix, E. and Hodges, J. L. (1951). Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties. Technical report, USAF School of Aviation Medicine. Technical Report 4, Project no. 21-29-004.

Hall, P. and Kang, K.-H. (2005). Bandwidth choice for nonparametric classification. *Ann. Statist.*, 33(1):284–306.

Okuno, A. and Shimodaira, H. (2020). Extrapolation Towards Imaginary 0-Nearest Neighbour and Its Improved Convergence Rate. *to appear* in NeurIPS2020.

Samworth, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist.*, 40(5):2733–2763.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York.