関数データに対する PU learning について

大阪大学 大学院基礎工学研究科 寺田 吉壱

2 値判別問題において，得られているデータに対するラベルが一部しか得られていない場合の判別問題は半教師付き判別問題と呼ばれている．半教師付き問題に対して提案されている多くの方法は，ラベル付きのデータが両方のクラス（正，負の両クラス）から得られている状況を想定している．一方で，正のクラスのデータの一部にしかラベルがついていない場合やラベルのついた負のデータが十分に得られていない場合，通常の半教師付き学習の方法を適用することができない．そこで，片方のクラスのデータの一部にしかラベルが得られていない状況を想定した 2 値判別問題に関する研究が，近年，機械学習分野において盛んに行なわれている．この半教師付き判別問題は PU learning (learning form only positive and unlabeled data) と呼ばれている．

まず，PU learning について概説する．ここでは，Elkan and Nato (2008) の枠組みを紹介する．(Y, R, X) を $\{\pm 1\} \times \{0, 1\} \times \mathcal{X}$ 上のある分布 $\mathbb{P}$ に従う確率変数とする．ここで，Y はクラスラベル，R はラベルの観測を表す応答変数（$R = 1$ ならばラベルが観測，$R = 0$ ならばラベルが欠測），X は特徴量である．また，$\pi := \mathbb{P}(Y = 1)$ とする．そして，次のようなメカニズムを考える．

$$\mathbb{P}(R = 1 \mid X, Y = -1) = 0, \quad \mathbb{P}(R = 1 \mid X, Y = 1) = \mathbb{P}(R = 1 \mid Y = 1) =: \lambda > 0$$

このとき，判別関数 $f : \mathcal{X} \to \{\pm 1\}$ に対する誤判別率 $R(f)$ は次のように表現できる．

$$R(f) = \frac{2 - \lambda}{\lambda} \lambda\pi R_+(f) + (1 - \lambda\pi) R_0(f) + \pi(\lambda - 1),$$

ここで，$R_+(f) := \mathbb{P}(f(X) \neq 1 \mid R = 1)$，$R_0(f) := \mathbb{P}(f(X) = 1 \mid R = 0)$ である．$\mathbb{P}(R = 1) = \lambda\pi$ であることに注意すると，$\lambda\pi R_+(f) + (1 - \lambda\pi) R_0(f)$ は R をラベルとみなした際の判別問題に対する誤判別率であるから，λ もしくは π の値が一致推定可能もしくは既知であれば，PU classification の問題は R に対する cost sensitive learning によって解くことができる．しかし，PU classification の枠組みにおいて λ または π の推定法は幾つか提案されているが，一般に一致推定は困難である．

本発表では，関数データにおける PU learning 問題に対して，関数データの潜在的な高次元性に着目することで，λ または π の推定を必要としない新しい方法を提案する．$\mathcal{I} \subset \mathbb{R}$ をある有界区間，$L_2(\mathcal{I}) := \{f : \mathcal{I} \to \mathbb{R} \mid \int_{\mathcal{I}} |f(t)|^2 \, dt < \infty\}$ とし，関数データが関数空間 $L_2(\mathcal{I})$ 上の確率変数として得られている状況を考える．適当な正則条件の下で，各クラスの関数データ X に対して，ある正の実数列 $\{\theta_s\}_{s=1}^{\infty}$ と基底関数列 $\{\varphi_s\}_{s=1}^{\infty}$ が存在し，Karhunen-Loéve 展開と呼ばれる以下の表現が得られる．

$$X(t) = \sum_{s=1}^{\infty} \{\mu_s + \sqrt{\theta_j} Z_s\} \varphi_s(t) =: \sum_{s=1}^{\infty} W_s \varphi_s(t) \quad (t \in \mathcal{I})$$

ここで，$\mu_s = \langle \mu, \varphi_s \rangle$，$Z_s$ は $\mathbb{E}[Z_s] = 0$ 及び $\mathbb{E}[Z_k Z_l] = \delta_{k,l}$ を満たす実数値確率変数である．この展開から，関数データが本質的には無限次元の確率変数 $\{W_s\}_{s \in \mathbb{N}}$ によって構成されていることがわかる．そのため，潜在的な関数データの無限次元性を引き出すことができれば，データの完全な分離が期待される．実際に，Delaigle and Hall (2012) では，通常の教師あり判別問題において漸近的に完璧な分類が達成可能な方法を提案している．そこで，Delaigle and Hall (2012) のような関数データの実数空間 $\mathbb{R}$ への射影がなぜ有効かを明確にし，射影に基づく関数データに対して有効な PU learning 法を提案する．そして，PU learning の枠組みにおいても，提案手法によりラベルなしデータの分類の誤判別率を漸近的に 0 にすることができることを示す．

参考文献

[1] Delaigle, A. and Hall, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B* **74** 267–286.

[2] Elkan, C. and Noto, K. (2008). Learning classifiers from only positive and unlabeled data. *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 213–220.