コストに基づく入力不確実性がある下でのレベルセット推定のための能動学習

理化学研究所　稲津 佑
名古屋工業大学, 理化学研究所, 物質材料研究機構　竹内 一郎

1. はじめに

近年, 様々な応用分野においてベイズ推測に基づいた未知関数の最適化のための能動学習が行われている. レベルセット推定 (LSE) は, 未知関数がある閾値 h よりも大きい領域と小さい領域を同定するものであり, 能動学習における重要なタスクのひとつである. 先行研究では, 未知関数の事前分布にガウス過程 (GP) を仮定し, これによって定まる信用区間を用いた効率的な LSE のための能動学習法が提案された. しかしながら, 先行研究では出力の不確実性は考慮していても, 入力の不確実性は考慮されていなかった. 本報告では, 入力に不確実性が伴い, かつ, 費やすコストに応じてその不確実性の度合いが変わる設定の下, GP に基づいた効率的な LSE のための能動学習法について述べる.

2. 設定

関数 $f : D \to \mathbb{R}$ を, $D \subset \mathbb{R}^d$ で定義された評価コストが高い black-box 関数とする. 各入力 $\boldsymbol{x} \in D$ に対し, 関数 $f(\boldsymbol{x})$ の値は $y = f(\boldsymbol{x}) + \varepsilon$ として観測されるとする. ただし, ε は正規分布 $\mathcal{N}(0, \sigma^2)$ に従う, 独立なノイズである. このとき, D の有限部分集合 Ω と閾値 $h \in \mathbb{R}$ に関し, f に対する上位集合および下位集合

$$H = \{\boldsymbol{x} \in \Omega \mid f(\boldsymbol{x}) > h\},\ L = \{\boldsymbol{x} \in \Omega \mid f(\boldsymbol{x}) \leq h\}$$

を同定することを考える. 更に, 本稿では, 入力に対しコストに依存した不確実性が伴う状況を考える. コスト $c_1, \ldots, c_k$ は, $0 < c_1 < c_2 < \cdots < c_k$ を満たすとする. 各コスト c_i, $i \in \{1, \ldots, k\} \equiv [k]$ と入力 $\boldsymbol{x} \in \Omega$ に対し, $\boldsymbol{x}$ を入力した際に実際に入力される値 $\boldsymbol{s}(\boldsymbol{x}, c_i)$ は確率変数 $\boldsymbol{S}(\boldsymbol{x}, c_i)$ からのランダム標本とする. ただし, $\boldsymbol{s}(\boldsymbol{x}, c_i) \in D$ かつ $\boldsymbol{S}(\boldsymbol{x}, c_i)$ は既知の密度関数 $g(\boldsymbol{s}|\boldsymbol{\theta}_{\boldsymbol{x}}^{(c_i)})$ を持つとする.

3. ガウス過程に基づくレベルセット推定

関数 f に対する事前分布に GP を仮定する. このとき, データ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^t$ が与えられた下での f の事後分布は再び GP となり, その事後平均と分散をそれぞれ $\mu_t(\boldsymbol{x})$ および $\sigma_t^2(\boldsymbol{x})$ とする. このとき, 各点 $\boldsymbol{x} \in \Omega$ に対し, 第 t 試行時における $f(\boldsymbol{x})$ に対する信用区間を $Q_t(\boldsymbol{x}) = [l_t(\boldsymbol{x}), u_t(\boldsymbol{x})]$ で定める. ただし, $l_t(\boldsymbol{x}) = \mu_t(\boldsymbol{x}) - \beta^{1/2}\sigma_t(\boldsymbol{x})$, $u_t(\boldsymbol{x}) = \mu_t(\boldsymbol{x}) + \beta^{1/2}\sigma_t(\boldsymbol{x})$ であり, $\beta^{1/2} \geq 0$ である. このとき, 精度パラメータ $\epsilon > 0$ を用いて, H および L を以下のように推定する:

$$H_t = \{\boldsymbol{x} \in \Omega \mid l_t(\boldsymbol{x}) > h - \epsilon\},\ L_t = \{\boldsymbol{x} \in \Omega \mid u_t(\boldsymbol{x}) < h + \epsilon\}.$$

4. 効率的な LSE のための能動学習

効率的な LSE を行うための, 次に評価すべき入力点および, その入力点の観測に費やすコストを決定するための獲得関数を与える. そこで, 「単位コストあたりの, 入力の不確実性を考慮した, 期待値的な分類増加量」に基づいた新しい獲得関数を提案する. 点 $\boldsymbol{s}^*$ を新たな入力点とし, $y^* = f(\boldsymbol{s}^*) + \varepsilon$ が得られたとする. 組 $(\boldsymbol{s}^*, y^*)$ が追加されたときの, H および L の推定集合を $H_t(\boldsymbol{s}^*, y^*)$, $L_t(\boldsymbol{s}^*, y^*)$ と書く. このとき, 入力点 $\boldsymbol{x} \in \Omega$ の観測にコスト c_i をかけたときの, 単位コストあたりの入力の不確実性を考慮した, 期待値的な分類増加量 $a_t(\boldsymbol{x}, c_i)$ は以下で与えられる:

$$a_t(\boldsymbol{x}, c_i) = c_i^{-1} \int \mathrm{E}_{y^*}[|H_t(\boldsymbol{s}^*, y^*) \cup L_t(\boldsymbol{s}^*, y^*)| - |H_t \cup L_t|] g(\boldsymbol{s}^*|\theta_{\boldsymbol{x}}^{(c_i)}) d\boldsymbol{s}^*.$$

本報告では, $a_t(\boldsymbol{x}, c_i)$ 最大化に基づいた能動学習法を提案する. 更に, いくつかの緩い仮定の下, 提案法は確率 1 で有限回の点の取得ですべての点の分類が終了する. 詳細については当日報告する.