ISS-A-055

ボトムアップクラスタリングを用いた ベイズ境界の近傍標本の探索

中村樹生† 蔭山昌幸† 千田将大† 片桐滋† 渡辺秀行†† 大崎美穂†
† 同志社大学 †† ATR

1. はじめに

最小分類誤り確率(ベイズ誤り)状態に対応するベイズ境界が持つべき性質(ベイズ境界性と呼ぶ)に基づく分類器学習法[1,2]の研究を行っている. この手法において特に複雑で時間のかかるクラス境界近傍標本の選択法の確立を目指し, ボトムアップ型のツリークラスタリングによって標本間の位置関係を調べる手法を検討する.

2. 境界近傍標本の選択

ベイズ境界性に基づく分類器学習法では, クラス境界付近にある境界近傍標本を用いて, その境界のベイズ境界性尺度を計算する. もしベイズ境界性尺度の計算に, 境界から離れた標本が紛れ込むと, 尺度の推定は真値から大きく偏る. 従って, この手法において, 境界近傍標本の適切な選択は重要な役割を持つ.

これまで, そうした境界近傍標本を求めるため, 標本空間において仮想的な境界上標本(アンカーと呼ぶ)を生成した上でその近くに境界近傍標本を求める手法[1]や 1 次元の誤分類尺度上で求める手法[2]が試されてきた. しかし, 前者はややアドホックな手続きを含む上に計算量が多く, 後者は高効率であるものの1次元空間への圧縮に伴う標本空間内の標本位置情報の欠損があるため, いずれもまだ最良の策とは言い切れない. そこで本研究では, 予め学習用標本どうしの近傍関係をツリー状に整理することで, 標本空間において境界近傍標本を効率的に選択する手法の確立を目指し, まずボトムアップ的に学習標本をクラスタリングし, それをツリー構造で整理する手法の構成を行う.

3. ボトムアップ型階層的クラスタリング

本研究で構築する, 単連結法を元にしたボトムアップ型階層的クラスタリングは, 以下の手順から成る.

1. 全学習標本のそれぞれ(対象標本)に関し, ユークリッド距離の意味で最近傍標本を求め, その対象標本とその最近傍標本との「対」を最小単位のクラスタとする.
2. 第 1 ステップにおいて, 全ての標本は1つ以上のクラスタに属し, 特に, 複数の「対」に同時に属する標本も存在し得る. その複数「対」をまたがる標本があるとき, 該当「対」を一つのクラスタに統合する. なお, この統合操作は, たどれる限りの「対」に対して繰り返す.
3. 第 2 ステップで得たクラスタ内でそのメンバ標本の平均(セントロイド)を計算し, それらの平均どうしが最近傍となるクラスタを新たなクラスタとして統合する.
4. 全標本が1つのクラスタに含まれるまで第 3 ステップを繰り返す.

4. 評価実験とまとめ

得られたクラスタの散布状況を確認するため, 2 クラス・2 次元の混合ガウス分布によって生成した合成データ(標本数:2,200)に, 構築したクラスタリング法を適用した結果を調査した(図 1). 特に境界近傍におけるクラスタの様子を把握するため, クラスタメンバ数が 55〜75 のクラスタで次式のベイズ境界性尺度値を計算し, その値の大小を赤色マーカの大きさとして表現し, 該当クラスタメンバ位置に上書きした.

$$\boldsymbol{H(x) = -\sum_{i=1}^{2} (P(C_i|x) \log_2 P(C_i|x))} \qquad (1)$$

ここで, クラスタメンバ数が小さ過ぎると尺度値推定の信頼度が低下し, 大き過ぎるとその偏りが大きくなる. 実験では, そうした場合の赤色マーカの散布状況も確認した. 適切なサイズのクラスタ選択は今後の課題であるものの, 図中, 赤マーカが 2 クラスの境界付近をほぼ正確に辿っている様子を確認できる.

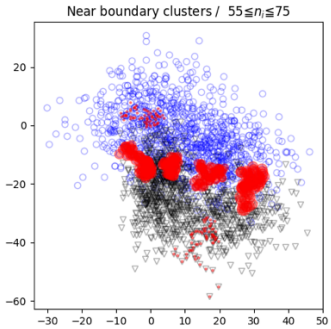

図1. 境界近傍のクラスタの散布状況の例. 青丸と黒三角はそれぞれ 2 クラスのうちの 1 つに属することを示している.

謝辞: 本研究の一部は, 科研費(18H03266)の支援を受けた.

参考文献

[1] Ha, D., et al. JSPS, Springer, 92, 135-151 (Feb 2020).
[2] Senda, M., et al. Proc. SPML, ACM (Nov 2019).