6R-07

最大ベイズ境界性学習法の実験的評価

大角朋弘† 片桐滋† 大崎美穂†
†同志社大学

1. 概要

パターン認識の分類器設計における究極の目的は最小分類誤り確率(ベイズ誤り)状態を達成することであり，その際の分類境界をベイズ境界と呼ぶ．ベイズ境界を選択する手法の一つとしてベイズ境界性に基づく分類器パラメータ選択手法(Bayes Boundary-ness-based Selection Method, BBS 法)[1]が提案されているが，時間的なコストが欠点としてあるため，この問題を解決する新しい手法であるベイズ境界性最大化学習法(Maximum Bayes Boundary-ness Training Method, MBB 学習法)[2]が近年提案された．この手法の目的は，学習を通して直接的にベイズ境界の達成を目指すものである．本稿では，MBB 学習法の有用性を検証するために評価実験を行う．

2. ベイズ境界性学習法(MBB 学習法)の概要

MBB 学習法は，次の 2 ステップで構成される． MBB 学習法は多クラスに適用可能であるが，以下では 2 クラスに適用した場合で考える．

・ステップ 1:ベイズ境界性尺度の設定(評価)
・ステップ 2:分類器パラメータの更新(学習)

ステップ 1 において，ベイズ境界尺度の式を以下のように定義する(式(1))．

$$H_v(x) = 0.42 - 0.5cos2\pi P(C_j|x) + 0.08cos4\pi P(C_j|x), \quad (1)$$

式(1)において，入力に分類事後確率を採用し，クラス間で等しい時に最大値 1，一方のクラスに偏った時に最小値 0 を得る．ベイズ境界性尺度では分類境界を含む小領域である境界近傍領域で推定が行われる(図 1)．

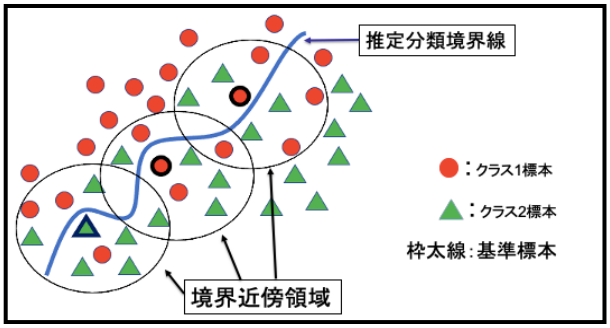


図 1. 境界近傍領域生成

Experimental Evaluation of Maximum Bayes Boundary-ness Training Method †Tomohiro Osumi, †Sigeru Katagiri, †Miho Ohsaki, †Doshisha University

図1において境界近傍領域は，推定分類境界に可能な限り近い標本を基準として k 近傍法を用いて生成された領域である．生成された境界近傍領域から分類事後確率を推定するために，標本に滑らかなカーネル関数を適用した擬似的な標本数を数える Soft-kNN 法を採用する(図 2)．

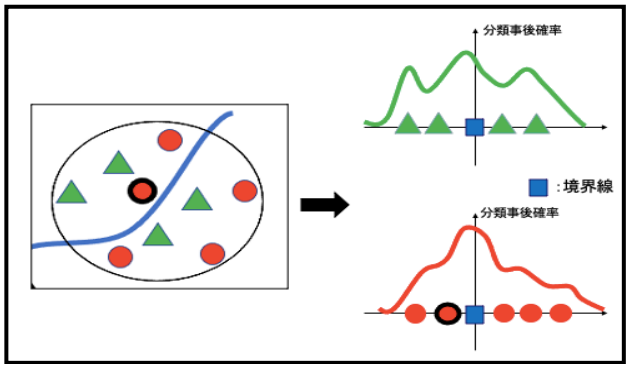


図2. Soft-kNN 法

図2において，高次元にある標本空間を分類境界までの距離的尺度を用いた1次元の誤分類尺度空間への写像を行う．ここで誤分類尺度とは，標本が帰属するクラスの帰属を示す識別関数値を用いる尺度である．識別関数値が各クラス等しい時，分類境界を表している．そして，写像された誤分類尺度空間上で Soft-kNN 法の標本個数の比率から分類事後確率の推定を行う．この分類事後確率を式(1)に用いることでベイズ境界性尺度の推定を行う．

ステップ 2 において，分類器パラメータの更新に際して，現状の分類器の状態と理想状態との差である損失を以下のように定義する(式(2))．

$$U_v(x;\Lambda) = 1 - H_v(x;\Lambda), \quad (2)$$

ただし，Λ は分類器パラメータを表している．

分類器パラメータの更新には最急降下法を用い，それによりステップ更新時に損失の減少が見込まれる．損失の減少はベイズ境界性尺度値の増加を意味するため，学習におけるベイズ境界性尺度の最大化の結果として目的であるベイズ境界の達成が期待される．
ステップ 2 の学習で分類境界が更新されるため，推定精度が低下する可能性がある，更新された分類境界に対してステップ1の推定を行う．以上の 2 ステップ(学習と評価)を繰り返すことで目的であるベイズ境界達成を目指す手法である．

3. 評価実験

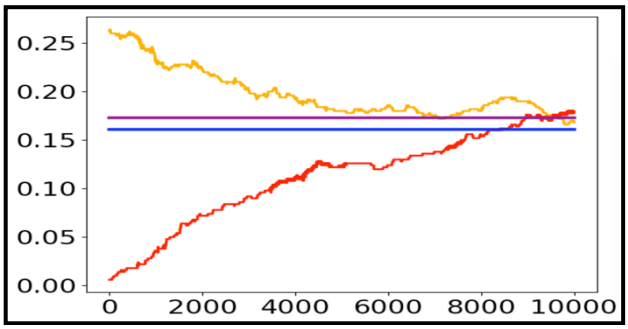
実験で使用数データセットを以下に示す(表 1).

表1 データセット一覧

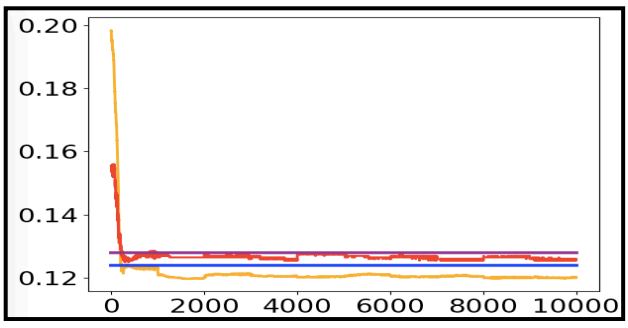
データセット	標本数	次元数	クラス数
GMM5C	1000	2	5
GMM2000	12000	2	2
Abalone	4177	7	3
Spambase	4601	57	2

表1において, GMM5C と GMM2000 は人工のデータセットで Abalone と Spambase は実世界のデータセットである. 各データセットは全標本を訓練用標本と訓練用標本に 1:1 で分割しており分類器の学習には学習用標本のみ適用している. 分類器についてはマルチプロトタイプ型(MPT)分類器を採用しており, プロトタイプ数は 10〜100 の 10 間隔で 10 パターンについて実験を行う. プロトタイプの初期化には k 平均法によるクラスタリングを使用する. 分類器パラメータの最適化手法は最急降下法とし, そのハイパーパラメータである学習係数は, 0.01, 0.05, 0.1, 0.5, 1.0, 5.0 の 6 パターンにおいて実験を行う. また, 学習回数を10000とし, ベイズ境界性尺度の再推定間隔を1000, 境界性近傍標本数を40とする. 最終的に採用される実験結果は, 全 60 パターンの中で損失の値が学習回数 10000 回の時に最も低い値を取ったものを採用する.

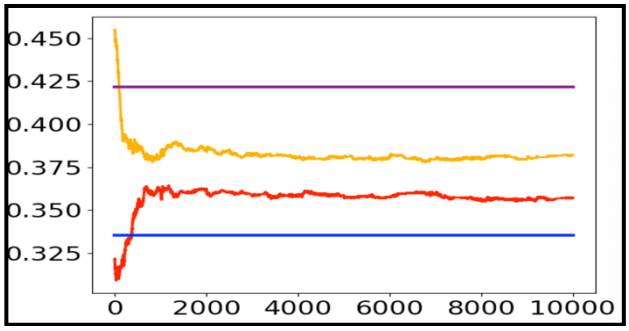
さらに, ベイズ誤りの参考値としてベイズ誤りを精度良く推定する交差検証法(CV 法)を採用する. 本実験では, Support Vector Machine (SVM)と MPT の CV 法(10-fold)を適用した場合を採用した.



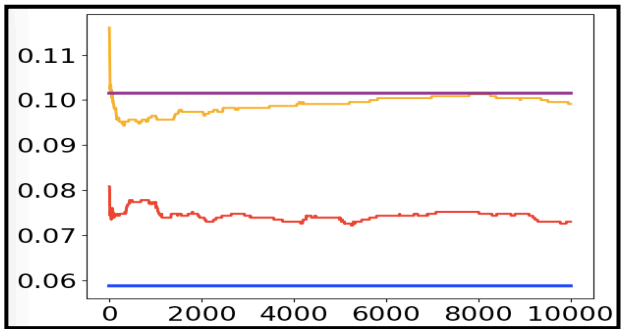
GMM5C



GMM2000



Abalone



Spambase

図 3. 実験結果

図 3 に, GMM5C と GMM2000, Abalone, Spambase の結果を示す. 横軸は学習回数, 縦軸は分類誤り率を表す. 図は, MBB 学習法の学習用と試験用標本の結果(赤, 黄)と, SVM 結果(青), MPT の CV 法の結果(紫)を示している.

4. 考察

図 3 から, MBB 学習法の試験用標本の分類誤り率(黄色の線)は, 比較手法の SVM と CV 法の分類誤り率である2本(青と紫)の線に漸近していることがわかる. 試験用標本に対する分類誤り率は, 未知の標本に対する誤り率に相当するものであるためベイズ誤りの推定値になる. このことから MBB 学習法は未知のベイズ誤りを直接的に推定しうることが示唆された.

5. まとめ

実験の結果から MBB 学習法はベイズ境界達成に対して有用性があることが示唆された. 今後の展望として MBB 学習法を他のデータセットに対して検証していく必要がある.

謝辞

本研究は科研費 18H03266 の支援を受けて行った.

参考文献

1) D. Ha, et al. JSPS, Springer (https://doi.org/10.1007/s11265-019-01451-y).
2) M. Senda, et al. Proc. SPML, ACM (https://doi.org/10.1145/3372806.3372817).