2P-04

正則化を用いた最大ベイズ境界性学習法について

松重　仁†　　片桐　滋†　　大崎　美穂†

同志社大学†

1　背景と目的

パターン認識における究極の目標は, 最小分類誤り率（ベイズ誤り）状態を達成する分類器パラメータの設定である. 損失最小化を経て直接的に分類器のベイズ境界性を高める新しい識別学習法, 最大ベイズ境界性（MBB: Maximum Bayes Boundary-ness）学習法[1)]が提案された. MBB学習法は入力標本ごとにベイズ境界性尺度を定義することで, 学習ステップ段階で評価でき, 分類器パラメータを最適化することで直接的にベイズ境界の達成を目指す. しかし, MBB学習法の評価は必ずしも十分に行われていない. 標本数が少ない時や標本が高次元で表される時に, 学習において過学習が起きている可能性があり比較手法である CV 法の分類誤り率の推定結果に十分漸近しない結果となっている. 本論文では, 以上のような問題点を解決するために, 分類器の学習に正則化を組み込み, 過学習の抑制を目指す.

2　最大ベイズ境界性学習法

2.1　概要

最大ベイズ境界性（MBB: Maximum Bayes Boundary-ness）学習法の学習目標に, ベイズ境界性尺度の最大化を採用する. ベイズ境界性とは, ベイズ境界状でクラス間の事後確率が等しくなることで分類判断が曖昧になるベイズ境界性の性質のことであり, その尺度のことをベイズ境界性尺度という. ベイズ境界性尺度が大きいほどベイズ境界に近づく.

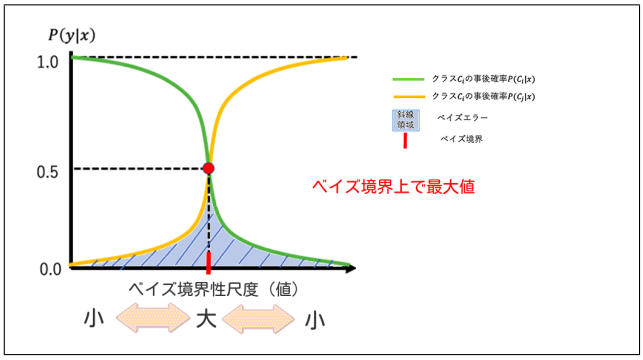

図1　2クラスの場合の分類事後確率とベイズ境界性尺度. 横軸は入力標本, 縦軸は各クラスの分類事後確率を表す.

Maximum Bayes Boundary-ness Training Method with Regularization
†Jin Matsushige †Shigeru Katagiri †Miho Ohsaki
†Doshisha University

MBB学習法は, 以下の2つのステップで構成されている.
Step1: ベイズ境界性尺度の推定
Step2: 分類器パラメータの更新

有限個の学習用標本上で計算される損失の最小化（ベイズ境界性の最大化）を目指す勾配法に沿う. まずStep1では, ベイズ境界性尺度値をシャノンエントロピー関数によって以下のように定義する.

$$H(x) = -\sum_{j=0}^{2} P(C_j|x)\log_2 P(C_j|x) \tag{1}$$

Step2の分類器パラメータの更新は, 最小化を目指す勾配法の一種である最急降下法を用いる. このStep2の学習の途中で, 分類器パラメータが変更されることにより, 分類境界が変更する. そのため, 一定数のエポックの繰り返し毎にベイズ境界性尺度の再推定を行う.

2.2　損失関数

学習によって得られた分類器パラメータから計算される値と理想的な学習目標との〝乖離の程度〟を表す損失関数を用いる. 損失が最小となるときベイズ境界性が大きくなる. シャノンエントロピー関数を用いて損失関数を以下のように定義する.

$$U_y(x;\Lambda) = 1 - H_y(x;\Lambda) \tag{2}$$

ただし, 右辺中の"1"は理想状態のベイズ境界性尺度を出力したものを表す. 式（2）は各入力標本に対する損失関数であるため, 分類器パラメータの更新はこの損失の境界近傍領域内の重み付き平均, すなわち以下の経験的平均損失の最小化を目指して行う.

$$L(\Lambda) = \sum_{x\in S_{NB}} W_y(x;\Lambda)U_y(x;\Lambda) \tag{3}$$

3　正則化を組み込んだ最大ベイズ境界性学習法

正則化項は, 入力された標本とベストインコレクトクラスの各プロトタイプとのユークリッド距離を用いた関数とし, 以下の式で表す.

$$g_{y^*}(x;\Lambda) = -\frac{1}{2}\|x - \lambda_{y^*}\|^2 \tag{4}$$

この正則化項を組み込んだ経験的平均損失として，以下のように定義する.

$$S(\Lambda) = \sum W_y(x;\Lambda)U_y(x;\Lambda) + \beta g_{y^*}(x;\Lambda) \tag{5}$$

ここで, β は正の定数である.

4 評価実験

4.1 概要

MBB学習法の有用性を検証するために, 今回の実験では4種類のデータセットを用いて評価実験を行った. ベイズ境界性尺度の最大化は, 分類境界のベイズ境界性を高めることになるので, 学習結果としてベイズ境界を満たすことを目的とする. ベイズ境界を満たしていることの検証については, ベイズ誤りを推定する手法の一つであるマルチプロトタイプ型分類器に対する交差検証（CV）法を用いた実験結果との比較を行う.

MBB学習法の有用性を検証するため4つのデータセットに対して検証実験を行う. 性質を以下の表に示す.

表1　データセット

データセット	学習用標本数	試験用標本数	次元数	クラス数
Abalone 01	1364	1366	7	2
GMM300	300	11700	2	2
German	500	500	24	2
WineRed3C	799	800	11	3

GMM300は人工のデータセットであり, 残り3つは実世界からから得た実データセット[3]である.

4.2 実験結果

以下に各データセットの実験結果を示す.

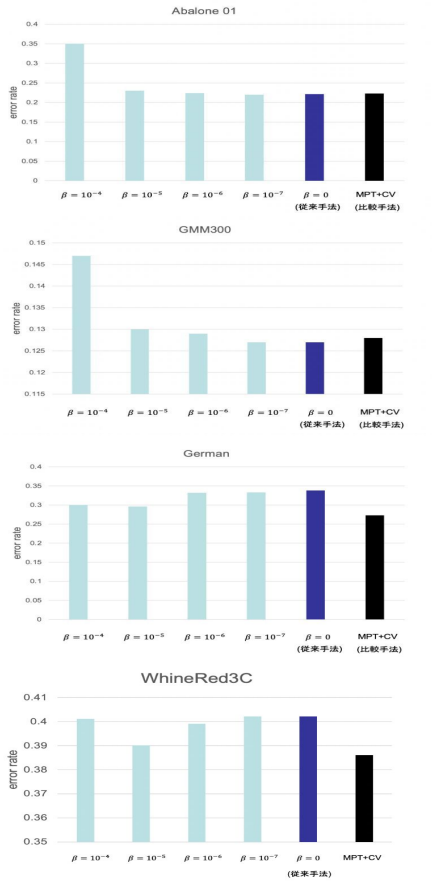

図2　上から, Abalone01, GMM300, German, WineRed3Cの実験結果. 縦軸を分類誤り率とし, 様々な正則化係数 β の値の結果と比較手法の結果を表している.

表2　実験結果（分類誤り率）

データセット	$\beta = 10^{-5}$	$\beta = 0$(従来手法)	CV法(比較手法)
Abalone 01	0.23	0.221	0.223
GMM300	0.13	0.127	0.128
German	0.296	0.338	0.273
WineRed3C	0.385	0.402	0.386

4.3 考察

正則化項を加える前からCV法を用いた比較手法の結果と近似しているGMM300とAbalone01では, 正則化係数の値が大きいと誤り率が大きくなり, 小さくなる（正則化項が0に近づく）と比較手法の値に近づく. つまり, 過学習が起きておらず正則化項を加えることで性能が少し悪くなったと考えられる. 対して, 比較手法の結果とあまり近似していなかったGermanとWineRed3Cでは, 従来手法の結果よりも誤り率が小さくなり比較手法の結果に近似するような正則化係数の値が存在した. 正則化項を加えることで過学習が抑制され性能が良くなったと考えられる.

正則化係数 $\beta = 10^{-5}$ の時にGermanとWineRed3Cでは誤り率が一番小さくなり, 比較手法の結果に近似するような結果が得られた. GMM300とAbalone01では誤り率が一番小さくはなっていないが, 比較手法の結果と近似していることが結果から分かる. また, データを扱う際に標準化をしているので正則化係数の値を $\beta = 10^{-5}$ とすると様々なデータセットに対しても適応でき, 正則化を組み込むことでMBBの有用性が確認できる.

謝辞

本研究の一部は科研費18H03266の支援を受けて行われた.

参考文献

1) Masahiro Senda, Ha David, Hideyuki Watanabe, Shigeru Katagiri, and Ohsaki Miho, "Maximum Bayes Boundaryness Training for Pattern Classification", SPML'19: Proceedings of the 2019 International Conference on Signal Processing and Machine Leaning, Association for Computing Machinery.
2) 千田将大. ベイズ境界性最大化学習法に関する研究. 同志社大学大学院理工学研究科. 修士論文
3) University of Carifornia, Irvine. UCI Machine Leaning Repository Retrieved September 25, 2019 from http://archive.ics.uni.edu/ml/index.php