Boosting を導入した Random Forests の提案

EP08056 鈴木 豊和　　指導教授：藤吉弘亘

1. はじめに

Random Forests(RF)[1] は複数の決定木を統合し，学習にランダム性を取り入れることで過学習による汎化性能の低下を防止したアンサンブル学習法である．RF は決定木の数やノード数の増加に伴い多くのメモリを必要とするため，小規模なハードウェア化には不向きであるという問題がある．そこで，本研究では Boosting[2] を導入した RF を提案する．逐次学習により相補的な識別器を構築することで，少数の決定木において識別性能の向上が期待できる．

2. 提案手法

提案手法では，誤識別した学習サンプルの重みを大きくすることで，次の学習ラウンドで誤識別したサンプルを識別できるように決定木を逐次的に構築する．提案手法の学習過程と識別過程を図 1 に示す．

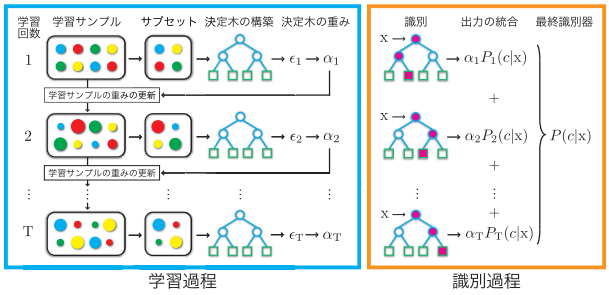

図 1：提案手法の流れ

学習過程

以下に，提案手法の学習過程のアルゴリズムを述べる．

Step1：前処理

N 個の学習サンプル $\{\mathrm{x}_i, y_i, w_i\}$ を用意する．ここで，$\mathrm{x} \in \mathbf{R}^d$ は d 次元のサンプルの特徴量，$y \in c$ は M 種類のクラスラベル，w は学習サンプルの重みを表す．学習サンプルの重み w は 1 で初期化する．

Step2：サブセットの作成

学習サンプルからランダムサンプリングによりサブセットを作成する．

Step3：決定木の構築

サブセットを用いて決定木を構築する．決定木はサンプル集合を分割する分岐ノードと各クラスに対する確率を有する末端ノードにより構成される．分岐ノードは分岐関数を用いて，ノードに到達したサンプル集合 S を左右の子ノードのサンプル集合である S_l と S_r に分割する．分岐関数はランダムに用意した特徴と閾値の組み合わせの中で，情報利得が最大となる組み合わせを選択する．各ノードの情報利得の算出において，重みの大きなサンプルを優先的に分割するため，クラス c_j の確率 $P(c_j|n)$ はサンプル i の重み w_i を用いて式 (1) のように算出する．

$$P(c_j|n) = \sum_{i \in S_n \wedge y_i = c_j} w_i \;/\; \sum_{i \in S_n} w_i \qquad (1)$$

ここで，S_n はノード n に到達したサンプル集合である．分岐を繰り返し，決定木の深さが一定の深さまで到達した場合，またはノードに到達したサンプル集合の情報利得が 0 の場合に末端ノードを作成する．末端ノードは式 (1) で求められるクラスに対する確率分布 $P(c|n)$ を保有する．

Step4：決定木の重み

決定木の重み α_t は決定木の誤識別率 ϵ_t から求める．決定木 h_t の誤識別率 ϵ_t は，学習サンプルの重み w_i を用いて式 (2) により算出する．

$$\epsilon_t = \sum_{i:y_i \neq h_t(x_i)}^{\mathrm{N}} w_{i,t} \;/\; \sum_{i=1}^{\mathrm{N}} w_{i,t} \qquad (2)$$

決定木の重み α_t を決定木の誤識別率 ϵ_t を用いて式 (3) により算出する．

$$\alpha_t = \frac{1}{2} \log \frac{(\mathrm{M}-1)(1-\epsilon_t)}{\epsilon_t} \qquad (3)$$

Step5：学習サンプルの重みの更新

学習サンプルの重み $w_{i,t}$ は決定木の重み α_t を用いて式 (4) により更新する．

$$w_{i,t+1} = \begin{cases} w_{i,t} \exp(\alpha_t) & \text{if } y_i \neq h_t(x_i) \\ w_{i,t} \exp(-\alpha_t) & \text{otherwise} \end{cases} \qquad (4)$$

学習サンプルの重みを更新後，サンプルの重みの総和が N となるように正規化する．**Step2** から **Step5** を T 回繰り返すことにより，T 個の決定木と決定木の重みを得る．ここで，決定木の重み α は合計値が 1 となるように正規化を行う．

識別過程

学習過程で作成した T 個の決定木の出力を統合し，入力 x に対する識別結果を出力する．最終識別器は，式 (5) のように各決定木の出力 $P_t(c|\mathrm{x})$ を決定木の重み α_t を用いて加重平均し，平均値が最大となるクラスを識別結果として出力する．

$$P(c|\mathrm{x}) = \sum_{t=1}^{\mathrm{T}} \alpha_t P_t(c|\mathrm{x}) \qquad (5)$$

3. 評価実験

提案手法の有効性を示すために，評価実験を行う．

3.1. 実験概要

評価実験は UCI Machine Learning Repository から 5 つデータセットを用いて誤識別率を比較する．実験に用いた学習パラメータは，決定木の数を 200，特徴選択回数をデータセットの特徴次元数の二乗根，閾値選択回数を 10，決定木の深さの上限を 7 とした．

3.2. 実験結果

提案手法と RF の各データセットの誤識別率を表 1 に示す．ここで，各手法の括弧内の数値は決定木の深さの上限値である．提案手法は従来手法の RF と比較して平均約 4%識別性能が向上した．

表 1：各データセットの誤識別率

手法	Pendigits	Letter	Wave	Satelite	Spambase	平均値
RF(7)	6.8	18.6	17.3	11.7	7.3	12.3
提案手法 (7)	**2.5**	**5.7**	17.0	**10.0**	**4.9**	**8.0**

データセット Pendigits における提案手法と RF の決定木の数に対する誤識別率を図 2 に示す．提案手法の誤識別率のカーブは従来手法の RF のカーブに対して左下にプロットされていることから，少数の決定木で高性能な識別を実現できたといえる．

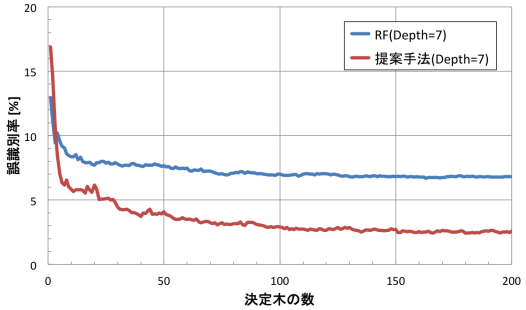

図 2：決定木の数に対する誤識別率

4. おわりに

本稿では，Random Forests に Boosting による逐次学習を導入することで，少数の決定木において識別性能を向上させることができた．今後は識別がより難しい画像認識問題において提案手法の有効性を検証する予定である．

参考文献

[1] L. Breiman, “Random Forests”, Machine Learning, vol45, pp.5-32, 2001.

[2] Y. Freund, R. Schapire, “Experiments with a new boosting algorith”, Machine Learning: ICML, pp.148-156, 1996.