学習データの被予測性能に着目した Alternating Decision Forests の各決定木への重み付け予測法

1X10C105-2　三沢翔太郎
指導教員　後藤正幸

1　研究背景と目的

近年，データマイニングの分野では様々な予測手法が提唱され，その有用性が示されている．それらの 1 つに特徴空間を分割することで予測ルールを構築する決定木が存在する．決定木では同じクラスの学習データを同一の特徴空間 (以下，ノード) に分割する作業を繰り返すことで木構造のモデルを学習する．しかし，決定木は学習データに依存してモデルが大きく変わり，汎化性能が安定しないという問題点がある．この問題は，ブートスラップサンプル等で学習した複数の決定木をアンサンブルすることで解決できることが知られている．この手法として Random Forests (以下，RF) や Alternating Decision Forests [1](以下，ADF) といった手法が既に提案されているが，本研究では特に予測精度の高い ADF に着目する．ADF は，個々の決定木を独立に生成するアンサンブル手法と比較し，モデル集合全体で学習データが分類されやすくなるよう，各決定木の同じ深さのノードを並列に分割していく．その際，各学習データにウェイトを付与し，これを活用してデータをより明確に分離するような分割法を採用していく．しかし，このウェイトは生成した各学習データの分類されやすさを表していると解釈できるものの，新規データを予測する際には活用されていない．

そこで本研究では，新規データの予測を行う際に学習時に算出したウェイトを用いて各決定木の出力を算出することで，予測精度が高いと想定される決定木の出力を重視して，予測精度の向上を図るアルゴリズムを提案する．また，ベンチマークデータを用いた実験を行い，提案手法の有効性を示す．

2　Alternating Decision Forests

2.1　問題設定

N 個の学習データの集合を $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$，離散カテゴリ集合を $\mathcal{Y} = \{c_1, c_2, \cdots, c_K\}$ とする．$\boldsymbol{x}_i$ は q 次元の説明変数ベクトルで，対応する目的変数 $y_i \in \mathcal{Y}$ は $\boldsymbol{x}_i$ に付与されているカテゴリとする．本研究では，$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ が得られたもとで，新たに $\boldsymbol{x}_{N+1}$ が与えられたとき対応する目的変数 y_{N+1} を予測する問題を考える．

2.2　概要

ADF は T 本の決定木を並列に生成し，それらをアンサンブルする予測モデルである．ADF の学習は，同時に全ての決定木の同じ深さのノードを，各ノードで分割基準となる変数をランダムに変えて分割する操作を指定の深さ D まで繰り返す．このとき，学習データを葉ノードまで到達させ，そこに存在する同じクラスの学習データの割合 (以下，純度) を求め，純度が高くなるように分割する．さらに，新たな分割を行う際には純度が低い学習データの純度が重点的に高くなるように分割するために，各学習データにウェイトを付与し利用する．このようにモデル集合を生成することで，各学習データの対応する葉ノードでの純度を全ての決定木で平均化した値を高くすることができる．新規データの予測時には，以下の式 (1) に従って予測クラスを推定する．

$$\hat{c}_k = \arg\max_{c_k} \frac{1}{T} \sum_{t=1}^{T} g_{m_{Dt}}(\boldsymbol{x}_{N+1}, c_k) \qquad (1)$$

ここで，深さ D の ADF を $\mathcal{M}_D$，その要素である決定木を $m_{Dt} \in \mathcal{M}_D$ とする $(1 < t < T)$．また，$g_{m_{Dt}}(\boldsymbol{x}_{N+1}, c_k)$ は決定木 m_{Dt} の説明変数 $\boldsymbol{x}_{N+1}$ に対応する葉ノードに存在するクラス c_k の学習データの割合を表す．ここで深さを 2 とした場合の ADF を図 1 で示す．

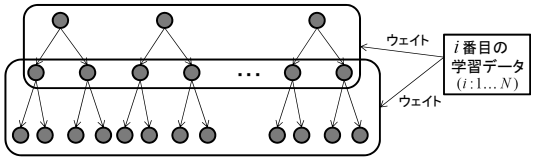

図 1．深さ 2 の ADF

2.3　ADF の学習

深さを $d = 1, 2, \cdots, D$ として，ADF の学習途中の深さ d の決定木集合を $\mathcal{M}_d$，その要素である各決定木を $m_{dt} \in \mathcal{M}_d$ で表す $(1 < t < T)$．ADF では学習途中の深さ d のモデル集合を用いた全ての学習データの純度の合計を損失関数で表し，これを極小化するように各決定木の葉ノードの分割を繰り返す．ここで深さ d のモデル集合に対する損失関数 L_d を式 (2) で表す．

$$L_d = \prod_{i=1}^{N} \exp\left\{-\frac{1}{T}\sum_{t=1}^{T} g_{m_{dt}}(\boldsymbol{x}_i, y_i)\right\} \qquad (2)$$

ここで，式 (2) は $g_{m_{dt}}(\boldsymbol{x}_i, y_i)$ の合計に対して単調減少である．すなわち L_d が小さいほど全ての学習データの純度が高くなり，学習データへの当てはまりが良くなっている．

ADF では，全ての学習データの純度を向上させるための分割方法選択を Gradient Boost [1] と呼ばれるアルゴリズムを用いる．このアルゴリズムでは，新たに分割したノードに所属する学習データの純度が高くなるように学習データにウェイトを付与し，その分割を行う．ここで，学習データ $(\boldsymbol{x}_i, y_i)$ のウェイトを $w_d(\boldsymbol{x}_i, y_i)$ として，式 (3) で定義する．

$$w_d(\boldsymbol{x}_i, y_i) = \exp\left\{-\frac{1}{T}\sum_{t=1}^{T} g_{m_{dt}}(\boldsymbol{x}_i, y_i)\right\} \qquad (3)$$

あるノードの分割を考える際，ランダムに選択した一部の変数集合を分割候補とする．それらの候補で分割したノードに存在する各クラスの学習データのウェイトの総和をそれぞれ計算し，それが最も偏るような変数を用いてノードを分割する．ここで，ウェイトを基準に分割方法を選択することで，ウェイトが大きい学習データの純度が改善する分割方法が選択されやすくなる．以上のウェイトの付与と分割の工程を指定の深さ D まで繰り返し学習を行う．

3　提案手法

3.1　従来手法の問題点と提案手法の概要

従来手法では学習データにウェイトを付与し，ADF 全体で損失関数が小さくなるようなモデル集合を構築する．ADF では，各学習データに対応する葉ノードでの純度を全ての決定木で平均化した値を高くすることのみを考えている．しかしながら，学習データのウェイトはノードの分割に使用するのみであり，新規データの予測時には使用しない．そこで本研究では，ウェイトを活用して新規データ予測時の各決定木

の出力を計算することを考える．具体的には，新規データ予測時に対応する葉ノードにある各クラスの学習データのウェイトを加味した出力を計算する．これにより学習時の情報を有効活用し，新規データの予測に用いるノードの純度が高く，予測の信頼性が高いと考えられる決定木の出力を重視した予測法を構築することで，予測精度の向上を図る．

3.2 ウェイトを活用した出力

提案手法ではウェイトで重みづけしたクラスごとの学習データの割合を個々の決定木で出力する．これにより，各葉ノードに存在する学習データのウェイトの平均が大きいクラスの出力が相対的に大きくなる．説明変数 $\boldsymbol{x}_i$ に対応する決定木 m_{Dt} の葉ノードを $s_{m_{Dt}}(\boldsymbol{x}_i)$ とし，決定木 m_{Dt} に新規データ $\boldsymbol{x}_{N+1}$ を入力した時のクラス c_k に関する出力を $f_{m_{Dt}}(\boldsymbol{x}_{N+1}, c_k)$ とすると，$f_{m_{Dt}}(\boldsymbol{x}_{N+1}, c_k)$ は葉ノードに存在する学習データのウェイトを活用して以下で定義される．

$$f_{m_{Dt}}(\boldsymbol{x}_{N+1}, c_k) = W(c_k | s_{m_{Dt}}(\boldsymbol{x}_{N+1})) \tag{4}$$

ただし，

$$W(c_k | s_{m_{Dt}}(\boldsymbol{x}_{N+1})) = \sum_{i=1}^{N} \frac{\gamma_{c_k}(y_i) \cdot \eta_{m_{Dt}}(\boldsymbol{x}_i, \boldsymbol{x}_{N+1}) \cdot w_D(\boldsymbol{x}_i, y_i)^{\alpha}}{\eta_{m_{Dt}}(\boldsymbol{x}_i, \boldsymbol{x}_{N+1}) \cdot w_D(\boldsymbol{x}_i, y_i)^{\alpha}} \tag{5}$$

$$\gamma_{c_k}(y_i) = \begin{cases} 1, & y_i = c_k \\ 0, & y_i \neq c_k \end{cases} \tag{6}$$

$$\eta_{m_{Dt}}(\boldsymbol{x}_i, \boldsymbol{x}_{N+1}) = \begin{cases} 1, & s_{m_{Dt}}(\boldsymbol{x}_i) = s_{m_{Dt}}(\boldsymbol{x}_{N+1}) \\ 0, & s_{m_{Dt}}(\boldsymbol{x}_i) \neq s_{m_{Dt}}(\boldsymbol{x}_{N+1}) \end{cases} \tag{7}$$

とする．また，$\alpha(>0)$ はウェイトの影響を制御するパラメータである．

ADF の出力は従来同様，個々の決定木の出力を平均化し，この値が最大であるクラスを式 (8) を用いて求め，これを新規データ $\boldsymbol{x}_{N+1}$ のクラスと推定する．

$$\hat{c}_k = \arg\max_{c_k} \frac{1}{T} \sum_{t=1}^{T} f_{m_{Dt}}(\boldsymbol{x}_{N+1}, c_k) \tag{8}$$

提案手法は新規データ予測時にウェイトを用いてクラスの所属度合を計算することで，クラス間の所属度合の偏りが小さい葉ノードの出力を一様に近づける．これにより，予測精度が高くないと想定される葉ノードからの出力の重みを相対的に下げる効果がある．

3.3 提案手法アルゴリズム

提案手法のアルゴリズムを以下に示す．

【学習フェーズ】

Step1) 全ての学習データに対して均等にウェイトを付与し，深さ $d = 1$ の決定木を T 本生成．

Step2) 全ての決定木に対し，深さ d のノードの分割を，ランダムに選択した一部の変数集合からウェイトを考慮して選択し，深さ $d+1$ の決定木を生成．

Step3) Step2 で生成された深さ $d+1$ の ADF に対する損失を式 (2) を用いて計算．

Step4) 式 (3) を用いて学習データのウェイトを更新．

Step5) Step2 へ戻る．このとき予め指定した深さ D に到達したら終了．

【予測フェーズ】

Step1) 各決定木を用いて新規データ $\boldsymbol{x}_{N+1}$ を葉ノードまで到達させ，式 (4) から出力を計算．

Step2) 各決定木の出力を平均化し，ADF の出力を計算．

Step3) 式 (8) を用いて新規データ $\boldsymbol{x}_{N+1}$ のクラスを予測．

□

4 実験

提案手法の有効性を示すために UCI 機械学習レポジトリを用いて実験を行う．比較対象として従来の RF と ADF を用いて，予測精度による比較を行う．

4.1 実験条件

実験では，公開データセット UCI 機械学習レポジトリのベンチマークデータセット 2 種類を用いた．データセットの概要を表 1 に示す．

表 1．データセットの概要

データセット名	クラス数	次元数	学習データ数	テストデータ数
Letter	26	16	16000	4000
ionosphere	2	34	281	70

評価指標はテストデータに対する平均誤り率とし，実験回数は従来同様 Letter は 5 回，ionosphere は 250 回行った．決定木の本数は $T = 100$ とし，個々の決定木の最大の深さを Letter は $D = 15$，ionosphere は $D = 10$ とする．また，予備実験から $\alpha = 3$ として提案手法の精度を示す．

4.2 実験結果及び考察

図 2 に実験結果を示す．図 2 より比較手法と比べ提案手法の予測精度が最も高くなった．また，Letter よりも ionosphere の方が提案手法の効果が大きかった．

提案手法ではウェイトを考慮した出力を行い，予測精度が高くないと想定される葉ノードからの出力の重みを相対的に下げた．このことにより，予測精度が高いと想定される葉ノードの影響が ADF の出力に対して相対的に強くなり，予測精度の高い出力が重視され，予測精度が向上したと考えられる．また，データセット Letter は ionosphere と比較して，学習終了の段階で既に学習データの純度が十分に高く，学習データ間のウェイトの差が小さくなり，提案手法の効果が出にくかったと考えられる．

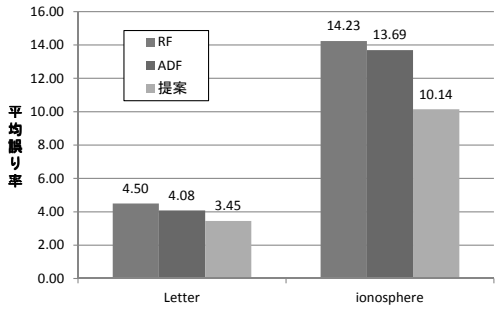

図 2．実験結果

5 まとめと今後の課題

本研究では，ADF を元に新規データ予測時にウェイトを活用した予測アルゴリズムを提案し，その有効性を示した．

今後の課題として，異なる損失関数を導入した提案手法の予測アルゴリズムの有効性や，提案手法の式 (5) に関して，データの構造と予測精度の関係や最適なパラメータ α の関係を検討することが挙げられる．

参考文献

[1]S.Schulter, P.Wohlhrt, C.Leistner, A.Saffari, P.Roth and H.Bischof, “Alternating Decision Forests,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp.508-515, 2013.