多重森林化による Global Refinement of Random Forest の予測精度向上法

1X12C076-7　土田知希
指導教員　後藤正幸

1　研究背景と目的

情報技術の発展により，データ分析に関する技術の重要性が増しており，その中でデータの所属するクラスを自動で予測する手法が特に必要とされている．これまでに様々な予測手法が提案されているが，本研究では Random Forests を改良した Global Refinement of Random Forest [1](以下，GRRF) と呼ばれる手法に着目する．GRRF は，決定木の集合からなる森全体で損失を最小化する手法 (Global refinement) と，重みが小さいノードを剪定してノード数を削減する手法（Global pruning) を組み合わせることで，Random Forests より高精度な予測を可能とする．

しかし，GRRF には 2 つの問題点がある．1 つ目は，大域的な最適化による重み算出の結果として重みが大きい一部のノードに予測が依存し，相関が低い複数の決定木を組み合わせる Random Forests の利点が損なわれるという点，2 つ目は，新規データが予測に対してあまり寄与しないノードに多く到達してしまった場合，その影響が強く現れて誤分類される可能性が高まるという点である．

本研究では GRRF の学習，予測フェーズに対し以下の改善を行うことで予測精度の向上を図る．学習フェーズでは森を任意の J 個の部分的な森 (以下，部分森) に分割することで (多重森林化)，相関が低い複数の決定木を組み合わせるという Random Forests の利点を維持する．予測フェーズでは重みが小さいノードを考慮しないことにより，重みが大きいノードを相対的に重視し，信頼性の高い予測を行う．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　Random Forests

N 個ある学習データ集合を $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$，離散クラス集合を $\mathcal{C} = \{c_1, c_2, ..., c_K\}$ とする．$\boldsymbol{x}_i$ は M 次元の説明変数ベクトル，$y_i \in \mathcal{C}$ は $\boldsymbol{x}_i$ に付与されているクラスとする．本研究では，$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ が与えられたもとで，新規データ $\boldsymbol{x}_{N+1}$ の所属するクラス y_{N+1} を予測する問題を考える．

Random Forests はブートストラップサンプルとランダムに選択された説明変数により相関が低い複数の決定木を独立に生成し，それらの出力結果の多数決により分類を行う手法である．いま，T 本の決定木を生成し，新規データ $\boldsymbol{x}_{N+1}$ の所属するクラス y_{N+1} を予測する場合を考える．まず，N 個の学習データからブートストラップサンプルを T 個生成する．続いて，各ブートストラップサンプルに対して全説明変数 M 個の中からランダムに選択した m 個 ($m < M$) の変数を用いて決定木を生成する．また，新規データ $\boldsymbol{x}_{N+1}$ のクラス y_{N+1} を予測する際は，新規データ $\boldsymbol{x}_{N+1}$ が到達したノードに含まれる学習データが最も多いクラスを決定木の出力とし，全決定木の出力の多数決により新規データ $\boldsymbol{x}_{N+1}$ のクラスを決定する．

3　Global Refinement of Random Forest

3.1　概要

Random Forest は独立に複数の決定木を生成し，その多数決により予測を行うため，森全体で最適化は行っていない．また，各決定木のノード数が多くなった場合，過学習が生じる恐れがある．GRRF は Global refinement と Global pruning を組み合わせることでこれらの問題点を解決する．Global refinement は森全体として最適な予測を行うため，予測される離散クラスである全決定木の葉ノードの出力を重みとしてその最適化を行う．Global pruning は重みが小さいノードを剪定してノード数を削減し，過学習の抑制を図る．Global refinement と Global pruning を交互に繰り返すことで森を作り，予測精度の向上を行う．

3.2　Global refinement

Global refinement では，森全体として最適な予測を行うために各学習データに対する出力とそのデータの実際のクラスとの差異により損失を定義し，その最小化を図る．このとき，各葉ノードからの出力を重みとし，それを森全体で最適化することにより実行される．なお，各葉ノードの重みは，そのノードの予測に対する寄与の大きさと解釈できる．また，森全体からの各学習データに対する出力は全ての葉ノードの重みから算出される．ここで，森の全葉ノード数をその次元数とするインジケータベクトル $\boldsymbol{\phi}(\boldsymbol{x}_i)$ と重みベクトル $\boldsymbol{w}$ を定義する．インジケータベクトル $\boldsymbol{\phi}(\boldsymbol{x}_i)$ は全決定木に対し，各学習データが到達した葉ノードに対応する要素を 1，それ以外の要素を 0 としたベクトル，重みベクトル $\boldsymbol{w}$ は各葉ノードの重みを要素とするベクトルとする．さらに，各学習データに対する出力を $\hat{y}_i$ とし，次式で定義する．

$$\hat{y}_i = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_i) \tag{1}$$

式 (1) における重み $\boldsymbol{w}$ を求めるため，以下の最適化問題を解く．

$$\underset{\boldsymbol{w}}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} l(\hat{y}_i, y_i) + \frac{\lambda}{2} \boldsymbol{w}^T \boldsymbol{w} \right\} \tag{2}$$

ただし，$l(\hat{y}_i, y_i)$ はヒンジロス関数であり，過学習を防ぐためにヒンジロス関数に正則化項 $\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$ を加える．なお，λ は正則化パラメータとする．式 (2) により全葉ノードの重みを変更し，森全体の損失の最小化が可能となり，精度の高い予測が可能となる．

また，新規データ $\boldsymbol{x}_{N+1}$ の所属するクラス y_{N+1} を予測する際は，式 (3) を基準に行う．ただし，多クラスの予測を行う際には，1 vs the rest 法を用いる．

$$\begin{aligned} \hat{y}_{N+1} = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_{N+1}) > 0 \Rightarrow \text{クラス } 1 \\ \hat{y}_{N+1} = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_{N+1}) < 0 \Rightarrow \text{クラス } -1 \end{aligned} \tag{3}$$

3.3　Global pruning

Global pruning では，重みが小さい，すなわち予測にあまり寄与しないと想定されるノードを剪定してノード数を削減し，過学習を防ぐことで予測精度の向上を目指す．同じ親ノードを持つ 2 つの葉ノードのペア全てに着目し，Global refinement で得られた重みがともに 0 に近い順に剪定を行う．このように重みが小さいノードを剪定することでノード数を削減して過学習を防ぎ，精度の高い予測を行う．

4　提案手法

4.1　提案手法の概要

GRRF には，新規データの予測の際は重みが大きい一部のノードに依存し，相関が低い複数の決定木を組み合わせる Random Forests の利点が損なわれているという問題点がある．また，重みが小さく，予測に対する信頼性が低いノードに新規データが多く到達してしまった場合，その影響が強く現れて誤分類される可能性が高まるという点でも改善の余地

がある．本研究では GRRF の学習，予測フェーズに対し以下の改善を行うことで予測精度の向上を図る．学習フェーズでは森を任意の J 個の部分森に均等に分割し，それぞれに GRRF を適用すること (多重森林化) を考える．これにより，相関が低い複数の決定木を組み合わせる利点を維持する．予測フェーズでは重みが小さいノードを考慮しないことにより，重みが大きいノードを相対的に重視し，信頼性の高い予測を行う．

4.2 多重森林化

GRRF では大域的な最適化を行うことで各ノードの重みを算出し，それを用いて予測を行うため，重みが大きい一部のノードに予測が依存し，相関が低い複数の決定木を組み合わせるという Random Forests の利点が損なわれてしまう．そこで提案手法では，GRRF の学習フェーズにおいて森を任意の J 個の部分森に分割し，それぞれの部分森の全葉ノードの重みをまとめたベクトル $\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_J$ を求める．それらを直列に結合することでベクトル $\boldsymbol{w}$ を求め，新規データ $\boldsymbol{x}_{N+1}$ が所属するクラス y_{N+1} の予測に用いる．これにより，全葉ノードに付加される重みのバラつきが減少し，相関が低い複数の決定木を組み合わせる利点が維持される．

4.3 重要なノードの考慮による予測

GRRF では Global pruning の実施後も重みが小さいノードが剪定されずに残されている場合がある．これは，ある葉ノードの重みが小さい場合でも，そのノードと同じ親ノードを持つもう一方のノードの重みが大きい，もしくは葉ノードではないために生じる．また，重みが小さく，予測に対する信頼性が低いノードに新規データが多く到達してしまった場合，その影響が強く現れて誤分類される可能性が高まる．そのため，提案手法では重みが大きい順に P% のノードのみで予測を行うこととする．これにより，重みが大きく，信頼性の高いノードを相対的に重視できるため，予測精度の向上が見込めると考えられる．

4.4 アルゴリズム

学習フェーズ

Step1) Random Forests と同様の手順で森を生成する．

Step2) 森を任意の J 個の部分森に分割する．

Step3) 各部分森に対し，式 (2) から $\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_J$ を求め，それらを直列に結合して $\boldsymbol{w}$ を求める．

Step4) 終了条件に達した場合は学習を終了する．そうでなければ **Step5** へ進む．

Step5) 同じ親ノードをもつ葉ノード全てに着目し，重みがともに 0 に近い順にあらかじめ定められた分だけ剪定を行う．

Step6) **Step3** に戻る．

予測フェーズ

Step1) 新規データ $\boldsymbol{x}_{N+1}$ が学習フェーズで生成された T 本の決定木で到達する葉ノードを発見する．

Step2) 新規データ $\boldsymbol{x}_{N+1}$ が到達した T 個の葉ノードのうち重みが大きい上位 P%の葉ノードを用いて所属するクラス $\hat{y}_{N+1}$ を求め，式 (3) を基準に分類を行う．□

5 実験

提案手法の予測精度面での有効性を示すため，ベンチマークデータを用いた実験を行う．比較対象として従来の GRRF を用いる．

5.1 実験条件

実験では UCI 機械学習レポジトリから 4 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 (サイズ)

データセット名	クラス数	説明変数	学習データ数	テストデータ数
ionosphere	2	34	281	70
breast-cancer	2	30	400	169
wine	3	13	100	78
Iris	3	4	90	60

評価指標は平均誤り率とし，実験はそれぞれ同じ条件で 50 回ずつ行った．決定木の本数 T は 100，個々の決定木の最大の深さは 14，正則化パラメータを $\lambda = 1$ とする．なお，予備実験により ionosphere では $J = 20, P = 50$，breast-cancer では $J = 5, P = 50$，wine では $J = 20, P = 70$，Iris では $J = 20, P = 90$ とした．

5.2 実験結果と考察

実験結果を図 1 に示す．その際，GRRF と提案手法で有意に差があるか t 検定を行う．図 1 における*は 5%有意，**は 1%有意を示している．

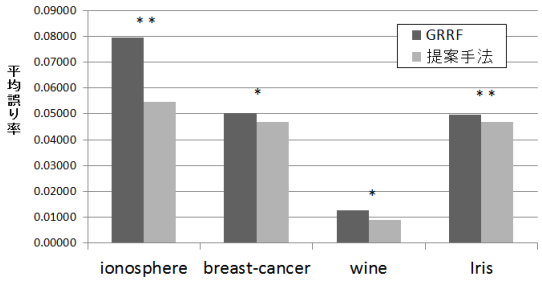

図 1. 実験結果

図 1 より，提案手法の予測精度は GRRF に比べて有意に高く，特に ionosphere と Iris では 1% 有意であることがわかる．これにより，提案手法が予測精度面で有効であるといえる．

提案手法では森を均等に分割して多重森林化を行い，それぞれを独立して最適化した．これにより，相関が低い複数の決定木を組み合わせる利点が維持されたため，予測精度が向上したと考えられる．また，提案手法では GRRF で剪定されずに残されていた重みが小さいノードを予測に用いずに分類を行った．これにより，重みが大きく信頼性が高いノードが相対的に重視され，予測精度が向上したと考えられる．さらに，データセット ionosphere は他のデータセットと比較して説明変数が多く，その中でも学習データが少ない．そのため，過学習が生じやすく，多重森林化によって過学習を抑制する効果が大きく現れ，予測精度が大きく向上したと考えられる．実際のデータは過学習が生じやすいため，提案手法の有効性が高いといえる．

6 まとめと今後の課題

本研究では，森を任意の J 個の部分森に分割して多重森林化を行い，それぞれ独立に最適化を行うことに加え，予測の際に重みが小さいノードを考慮しないことによって GRRF の予測精度を向上させる手法を提案し，実験を通じてその有効性を示した．

今後の課題として，部分森間の具体的な関係性の考察や，森の分割数 J や考慮するノードの割合 P を含む適切なパラメータ設定の決定法などが挙げられる．

参考文献

[1] S.Ren, X.Cao, Y.Wei and J.Sun, “Global Refinement of Random Forest,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp.723-730, 2015.