6R-4

重みつきサンプリングによるランダムフォレスト法

川久保　秀子†　　吉田　裕亮†
†お茶の水女子大学　大学院人間文化創成科学研究科

1　はじめに

変数選択は、全変数の中から意味のある変数を選択する際に用いる手法で，機械学習や統計学では特徴選択や属性選択，バイオインフォマティクスでは遺伝子選択などと呼ばれている．変数選択を行うことにより，次元の呪いを緩和して計算量を減少させたり，ノイズ変数を排して予測精度を向上させたり，複雑なデータ構造を要約して解釈し易くしたりすることができる．それ故，多くの分野で大規模データにおける変数選択は重要な役割を持ち，ランダムフォレスト法は変数選択の実用的な手法として用いられてきた．本研究では，既存のランダムフォレスト法を拡張し，大規模データの効率的な変数選択法を提案する．

2　ランダムフォレスト法

アンサンブル学習であるランダムフォレスト法 [1] は、決定木 CART を弱学習器として用い，相関の低い決定木群を作成することで，分類や回帰，変数選択を行う方法である．

2.1　決定木 CART

決定木 CART [2] は不純度が最も高い変数を選んで分岐を行う学習器である。不純度の計算過程で算出される Gini 係数が変数の重要度の指標となることから，ランダムフォレスト法では決定木 CART から算出された Gini 係数を用いて変数選択を行う．不純度は変数を分岐する前と後とでの誤差の改善度を表し，以下のように定義される．

$$\Delta GI(t) = P_t GI(t) - P_L GI(t_L) - P_R GI(t_R) \quad (1)$$

$GI(t)$ は Gini 係数と呼ばれ，以下のように定義される．

$$GI(t) = 1 - \sum_k p(k \mid t)^2 \quad (2)$$

$p(k \mid t)$ はノード t 内のクラス k が正しく分類されている比率，$GI(t_L)$ はノードの左側の枝の Gini 係数，$GI(t_R)$ はノードの右側の枝の Gini 係数，P_t は分割する前のサンプル数の比率，P_L は分割した後の左側のサンプルの比率，P_R は分割した後の右側のサンプルの比率を表す．

LaTeX 2_ε Style File for National Convention of IPSJ
†Graduate School of Humanities and Sciences, Ochanomizu University

決定木 CART では全ての組み合わせを調べ，木を成長させる．

2.2　ランダムフォレスト法のアルゴリズム

ランダムフォレスト法のアルゴリズムのうち，本研究に関係する部分を以下に示す．ここではサンプル数を n, 次元数 (全変数の数) を M とする．

1. N 組のブートストラップサンプル $B_i(i = 1, \cdots, N)$ を復元抽出法で作成する．
2. ブートストラップサンプル $B_i(i = 1, \cdots, N)$ における M 個の変数の中から m 個の変数を一様分布に従いランダムサンプリングする．
3. ブートストラップサンプル B_i を用いて決定木 CART を生成する．ただし，生成した決定木 CART には剪定を行わない．
4. 決定木 CART の生成過程で算出した全ての Gini 係数の平均を取り，変数の重要度を求める．Gini 係数が大きければ変数の重要度は高くなる．

3　提案手法

Genuer らは，ランダムフォレスト法における変数のランダムサンプリング時に重みづけを行うアイデアを示唆した [3]. 本研究では具体的にどのように重みづけを行うかを考察し，ランダムフォレスト法の拡張を以下のように行うことを提案する．

1. I 組のブートストラップサンプル $B_i(i = 1, \cdots, I)$ を復元抽出法で作成し，予備推定を行う．ただし，決定木 CART を生成する際に不純度の計算を第 1 分岐点で止める．重要度の高い変数上位 5 位までに対し順位を反映したスコアを与え，生成した I 本分のスコアを集計する．決定木 CART では各分岐点において不純度の計算を最大で $2^{n-1} - 1$ 回行うので，計算を第 1 分岐点で止めることにより，計算量を減少させることができる．
2. 予備推定によって得られたスコアを基に，各変数に対して重みづけを行う．ここで，ギブス分布の

確率関数を定義する．ただし $\beta > 0$ とする．

$$P_i = \frac{\exp(-\beta G_i)}{\sum_{i=1}^{M} \exp(-\beta G_i)} \qquad (i = 1, \cdots, M) \quad (3)$$

予備推定で得られたスコアを正規化した値 G_i をポテンシャルとしてギブス分布を求める．

3. 新たにブートストラップサンプル $B_j (j = 1, \cdots, J)$ を復元抽出法により作成し，本推定を行う．生成する決定木 CART の総数が減少すれば，ほぼ線形に総計算時間が減少するため，ブートストラップサンプルの数は $N \geq I + J$ とする．m 個の変数のサンプリングは β の値を適当にチューニングしたギブス分布に従って行う．このサンプリングでは予備推定で重要度の高かった変数が選ばれ易くなっており，β の値を大きくすると重要度が高い少数の変数を選んで返す傾向がある．

4. ブートストラップサンプル B_j に対して決定木 CART を作成し，予備推定と同様に不純度の計算を第 1 分岐点で止め，重要度の高い変数を調べる．最も重要度の高かった変数にスコアを与えるが，この変数がギブス分布によって重複回選ばれている場合は重複して選ばれた回数をスコアとし，J 本分のスコアを集計する．

以上により，β の値を調整することで変数の重要度順に変数選択が行われるようになる．

4 実験

UCI repository の Internet Advertisements 及び，NIPS 2003 variable selection benchmark の GISETTE, ARCENE, MADELON をデータセットとして用いる．これらのデータセットのクラス数は 2 である．ブートストラップサンプル B_i，B_j に対して $m = \lfloor \sqrt{M} + 0.5 \rfloor$ 個の変数を提案手法に従って重みつきサンプリングし，I,J の値をそれぞれ 100 以下として実験を行う．各データセットのサンプル数を n，次元数を M，提案手法によって選択された変数の数を k とし，k のみを用いて SVM で求めた正答率 (選択後) と，全変数を用いて SVM で求めた正答率 (全変数) を表 1 に示す．また，図 1 には，$N = I + J = 100$ として，既存手法と提案手法の各過程で決定木 CART が要する計算時間を示す．なお，計算環境は CPU Phenom X4 9950，OS Windows7 Professional 64bit，RAM 8GB である．

4.1 結果

k と正答率 (選択) には，少ない変数で高い正答率が得られた際の結果を示した．表 1 から，変数選択後も

表 1: データセット別 変数の数と正答率

データセット	n	M	k	正答率 (選択後)	正答率 (全変数)
Internet	3,279	1,558	10	96.3	96.3
GISETTE	6,000	5,000	45	96.4	96.4
ARCENE	100	10,000	8	79.0	61.8
MADELON	2,000	500	28	56.0	57.9

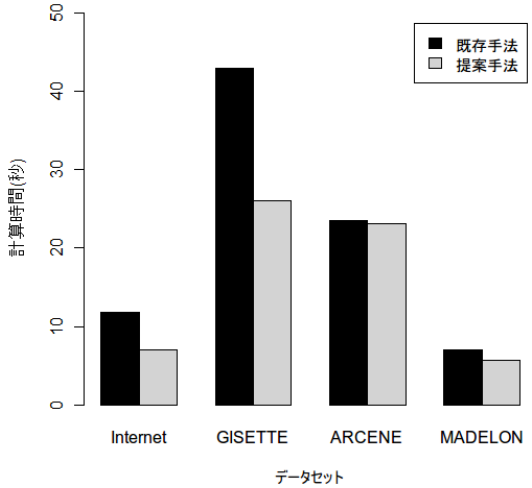

図 1: 各手法の過程で決定木 CART が要する計算時間

分類に必要な情報が維持され，適切な変数選択が行われたことが確認された．また，図 1 からは，計算時間が短縮されたことが確認された．ARCENE では計算時間の差が現れなかったのは，他のデータセットの場合と異なり，分岐数 4 以下の決定木 CART がほぼ 9 割を占めたことが原因として考えられる．

5 まとめ

ギブス分布を用いた重みつきサンプリングを導入することによって，既存のランダムフォレスト法を拡張し，大規模データにおける変数選択を効率的に行うことができた．今後の課題として，選択したい変数の数が決まっている時，β のチューニングを効率的に行う方法を考察していきたいと考えている．

参考文献

[1] Breiman, L.: Random forests, *Machine learning*, **45**, 5–32, (2001).

[2] Breiman, L.: *Classification and regression trees*, Chapman & Hall/CRC (1984).

[3] Genuer, R., Poggi, J.M., and Tuleau, C. *Random Forests: some methodological insights*, Arxiv preprint arXiv:0811.3619 (2008).