Vol.2011-MPS-86 No.33
Vol.2011-BIO-27 No.33
2011/12/2

重みつきランダムサンプリングによるランダムフォレスト法

川 久 保 秀 子[†1]

ランダムフォレスト法は，変数選択に用いることのできる方法の 1 つである．本研究では既存のランダムフォレスト法を拡張し，重みつきランダムサンプリングを導入することによって変数選択を効率的に行う方法を提案する．

Random Forest by a weighted random sampling

Hideko Kawakubo[†1]

Random Forest is one of the methods of using for variable selection. In this study, we improve Random Forest and suggest a certain efficient method for variable selection by introducing weighted random samplings.

1. はじめに

変数選択は機械学習においては特徴選択とも呼ばれ，説明変数の中からノイズ変数を排して意味のある変数を選択することにより，計算量を小さくするためなどに用いられている．ランダムフォレスト法 [1] はアンサンブル学習の 1 つで，変数選択に用いることが可能である．ランダムフォレスト法は高速でかつ精度が高いため，特に大規模データのデータマイニングに適した方法であるが，データの次元数が非常に高い場合には大きな計算量を要することがある．そこで本研究では，予備推定を行った後に変数の重要度からなる分布に従って変数のランダムサンプリングを行うことにより，効率的に変数選択を行う方法を提案する．

†1 お茶の水女子大学大学院　人間文化創成科学研究科
Graduate School of Humanities and Sciences, Ochanomizu University

2. ランダムフォレスト法

ランダムフォレスト法は，多数の決定木を用いたアンサンブル学習で，ランダムサンプリングされたトレーニングデータに対して説明変数をランダムにサンプリングし，相関の低い決定木群を作成する方法である．

2.1 決定木 CART

アンサンブル学習によく用いられる決定木として CART [2] が挙げられる．CART では分岐する変数を選択する際に不純度または情報量を用いる．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．

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

$GI(t)$ はノード t における Gini 係数と呼ばれ，以下のように定義される．

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

$p(k \mid t)$ 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．CART では不純度が最も高い変数を選んで分岐を行い，木を拡張させる．

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を以下に示す. データは M 次元とする．

(1) データの約 1/3 をテストデータとして除く．これを OOB データと呼ぶ．残りをトレーニングデータとし，I 組のブートストラップサンプル $B_1, \cdots, B_i, \cdots, B_I$ を復元抽出法で作成する．

(2) ブートストラップサンプル $B_i (i = 1, \cdots, I)$ 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．ヒューリスティックな値としては，$m = \sqrt{M}$ がよく用いられる．

(3) ブートストラップサンプル B_i に対し CART の Gini 係数や ID3，C4.5 のインフォメーションゲインなどの規準を用いて最良の特徴を選ぶ．ただし，生成した決定木には剪定を行わず，各ノードに分類されるクラスが 1 になるまで木の拡張を続ける．

(4) 生成された I 本の木それぞれに対して OOB データを用いてテストを行い，推定誤差を求める．これを OOB 推定誤差と呼ぶ．分類問題では多数決，回帰問題では平均をとり，その結果によって更に決定木の数を増加させ，新たに分類器を構築する．

Vol.2011-MPS-86 No.33
Vol.2011-BIO-27 No.33
2011/12/2

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) 初めに既存のランダムフォレスト法により予備推定を行う．この予備推定によって得られた変数の重要度を基に各変数に対して重みづけを行う．

(2) ギブス分布の確率密度関数を以下に定義する．

$$P_i = \frac{\exp(-\beta G_i)}{\Sigma_{i=1}^{M} \exp(-\beta G_i)} \tag{3}$$

変数の重要度を正規化した値 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 $B_j(j = 1, \cdots, J)$ を作成し，β の値を適当にチューニングしたギブス分布に従って m 個の変数をサンプリングする．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少数の重要度の高い変数を選んで返す傾向がある．

(4) 本推定としてブートストラップサンプル B_j に対して決定木を作成し，変数の重要度を計算する．最も大きな重要度を持つ変数に対して，その重要度の値を重複して選ばれた回数倍したものをスコアとし，スコアが 0 より大きい変数を選択する．

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. 実験例

UCI repository の Wisconsin Breast Cancer データセットを用いて実験を行う．このデータはクラス数 2，次元数 30，サンプル数 569 である．決定木は CART，分岐する変数を選択する際には不純度，変数の重要度には予備推定で計算された全決定木の Gini 係数の平均値を用いる．$m = \sqrt{M}$ 個の変数を重みつきランダムサンプリングし，I，J の値をそれぞれ変えて実験を行う．提案手法によって変数選択された変数のみを取り出し，$n = 10$ として n 重交差確認法により SVM で正答率を求める．更に OOB データを変えて提案手法の手順を最初から繰り返し，平均正答率を求める．この実験では，分類に必要な情報が変数選択後も維持されているかどうかの基準の正答率として，30 変数全てを用いて SVM で交差確認を行った時の平均正答率 0.964 を比較に用いる．また，図 1 から既存のランダムフォレスト法では安定した結果を得るために 400 程度の決定木が必要であることがわかるので，

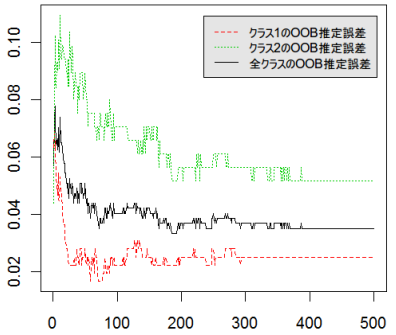

図 1 決定木の数と OOB 推定誤差

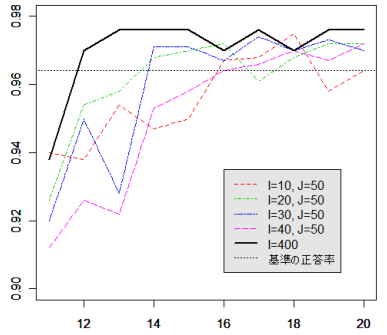

図 2 変数選択後の平均正答率

$I = 400$ として既存の方法でも変数選択を行い，比較に用いた．

4.1 結果

図 2 からは変数を 16 まで減少させても基準の正答率 0.964 と $I = 400$ の間に実験値がほぼ入ることがわかり，既存の方法よりも少ない計算量で適切な変数選択が行われたことが確認された．また，変数選択する変数の数 S を既知とし，降順にソートした P_i の上位 S 個の和を 0.9 以上にする β の値を $\beta_{0.9}^{S}$，n を全サンプル数とするとき，$\beta = j\beta_{0.9}^{S}\sqrt{i}/n$ で，選択したい数だけ変数を選択する確率の高いギブス分布が得られることがわかった．

5. まとめ

ギブス分布を用いた重みつきランダム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，変数選択を効率的に行えることが確認された．今後の課題として，全説明変数に対して意味のある説明変数が極端に少ない場合においても変数選択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).