ランダムフォレストによるクラス分類

富山県立大学電子・情報工学科　　指導教員：奥原浩之
1515028　杉山桃香

1　はじめに

今、分類や回帰の問題を扱う場合、選択する手法としてサポートベクターマシン (SVM) やランダムフォレストなどの様々な手法が候補に挙げられる.

そこで、今回はランダムフォレストという手法に着目した. ランダムフォレストとはどのような学習アルゴリズムでどのような利点・欠点があるかを調べ、その特徴を理解することを今回の目的とする.

更に、従来の手法と比較し、それぞれの手法の向き不向きとランダムフォレストが従来の手法よりも優れている点について考える。

2　ランダムフォレストの概要

ランダムフォレストは、集団アルゴリズムの一つである.1996 年に Breiman によって提案された. データ集合からランダムに抜き出したサンプルをもとに二分決定木を複数本作る. これを用いて、学習器を構成する. これは、クラス推定にも回帰問題にも用いることができる.

3　ランダムフォレストの仕組み

3.1　決定木の仕組み

ランダムフォレストを理解するためには、決定木学習の手法について理解する必要がある. よって、最初に決定木学習の理論について説明する.

決定木は親から順に条件分岐を辿っていくことで、結果を得る手法である. 図 1 に決定木のイメージを示す.

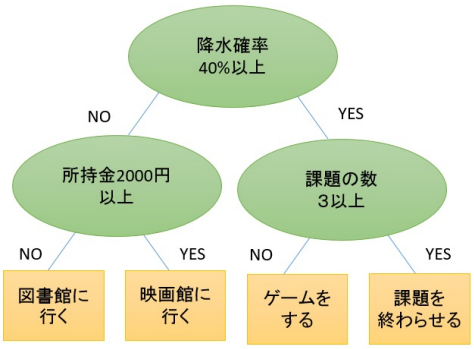

図 1　決定木のイメージ

決定木学習とはデータの応じて上の図のような決定木を構成し、分類を行う機械学習の手法のことを指す. 決定木学習は、データの種類に応じて決定木を成長させていく. 決定木の分類条件は、データを分類したときの情報 IG(Infomation Gain) が最大になるようにすること. 情報利得は式 (1) で表される.

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \tag{1}$$

Dp は親のデータ、N はノード、j は注目しているデータを表す. m は木を分割するノード数である. 一般的に決定木は二分木として実装されるので、ほとんどの場合は m=2 となる.

I は不純度という指標で、含まれるデータに偏りがあるほど大きな値になる. 不純度の計算にはエントロピー、ジニ不純度などが用いられる. 今回はエントロピーを使って説明する. 式 (2) にエントロピーの式を示す. ゲイン

$$I_H(t) = -\sum_{i=1}^{c} p(i|t) \log_2 p(i|t) \tag{2}$$

p(i—t) はデータ数 t の中に含まれているデータ数 i の割合を表す.

エントロピーはデータのばらつきが大きいほど大きな値となる. 例えばサンプル数 100 個のデータを分類したとき、左の木に 100 個、右の木に 0 個に分類すると、エントロピーは最も大きな値である 1 になる. 逆にエントロピーが最も小さくなるのは、左右の木に 50 個ずつ分類したときで、エントロピーは 0 になる.

式 (1) の I にエントロピー Ih を代入することで情報ゲインの計算を行う.

3.2　なぜ集団学習するのか

集団学習とはたくさんの弱学習器の結果をまとめ合わせ 1 つの識別器を構築する学習方法である. ランダムフォレストに限らず、弱学習器では集団学習が有効な場合が多い. 何故集団学習するのかの答えを考えるにあたって、まず、汎化誤差というものについて知っておく必要がある.

汎化誤差とは、バイアス (学習モデルの単純さに由来する誤差) とバリアンス (学習データの違いに由来する誤差) と削除不能な誤差の和である.

線形のような単純なモデルでは、バイアスが大きくバリアンスが小さい. よって、単純なモデルであるため、ノイズに強いが複雑な表現は出来ない.SVM や最小二乗法はこのタイプである.

一方、高次の複雑なモデルでは、バイアスは小さいがバリアンスは大きい. よって、複雑な表現が可能だが、過学習してしまいノイズも再現してしまう. ニューラルネットや決定木はこのタイプである.

汎化誤差について理解した上で、もう一度なぜ集団学習をする必要があるかについて考える. その答えは、集団学習を行うことによって、バリアンスを下げることが可能だからだ.

3.3　他の集団学習との違い

ランダムフォレストの他にもいくつか集団学習があるが、他の集団学習と何が違うのかについて考える. 集団学習の 1 つにバギングという手法があるこれは、全教師入力データからランダムにとったデータで、複数の学習器を作成する。

対して、ランダムフォレストは全教師入力データからランダムにとったデータで、複数の学習器の作成を行い、更に説明変数もランダムに抽出する. 説明変数もランダムに取って来る点が他の集団学習とは違う.

説明変数同士に相関があると、弱学習器間の相関が生まれる. よって、弱学習器間に相関があるとバリアンスが下がらない. ランダムフォレストのように、説明変数もランダムに選ぶため相関の低い決定木群を作成しバリアンスを下げることができる.

3.3　処理の流れ

[1] サンプルデータ集合から、ランダムにデータを選択してサンプル集合とする. 重複や使われないデータがあっても構わない.

[2] 1 を繰り返してサンプル集合を n 個作る.

[3] それぞれのサンプル集合の変数から 2 分決定木の根ノードを n 本作る.

[4] 決定木の各ノードを分岐関数 h(v, θ) で、サンプルを 2 つに分割し、2 つのノードを作成する.

[5] それぞれの木で、終了条件を満たすまで再帰的にノードを作り続ける.

2 分木の各ノードの分岐関数 h(v, θ) は単純な線形関数で、2 クラス分類を行うため出力は 0 もしくは 1 である. パラメータ θ はパラメータ候補の中から選ぶ. しかし, すべての候補を試すと計算量が非常に多くなるため、パラメータ候補はランダムに抽出したものを利用する.

1 つのクラスとそれ以外への分割を繰り返すことで、最終的な出力を全ての決定木で求め、多数決で結果を出す. イメージを図 2 に示す.

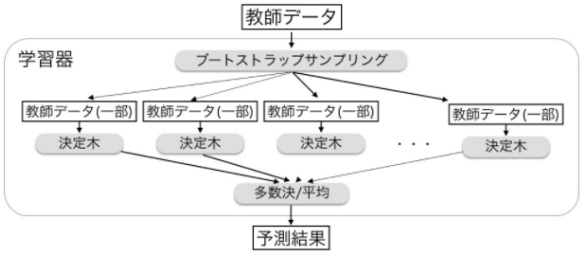

図 2 ランダムフォレストの仕組み

3.4 説明変数の重要度

特徴量の重要度の説明に入る前に、まず、重要度に関係するランダムフォレストの性質について説明する. ランダムフォレストでは、各決定木で異なるサンプルを使って学習を行う.

これを実現するために、学習データをそのデータ数だけ重複サンプリング（ブートストラップ法）し、それを学習サンプルとして決定木を学習させる. このとき、学習データのうち、平均で 1/3 ぐらいは学習に使われないデータがある. この使われなかったデータを out-of-bag（OOB）という.

次に、この学習の過程で得られた OOB を特徴量の重要度に用いることを考える。基本的なアイデアは OOB の各サンプルの値を混ぜて、その混ぜたサンプルで推定したら、どれ位精度が下がってしまうのかを調べる.

具体的な説明をするために、まず、ある学習済みの決定木について考える. その決定木の OOB を決定木で分類すると、OOB の各サンプルのうち、間違って分類されてしまったサンプルが現れる. この間違って分類されてしまった率を、OOB における誤り率と呼ぶ.

この OOB における誤り率を使い、特徴量の重要度を計算する. 特徴量の中のある変数を指定し、その変数の値を OOB の各サンプル間でランダムに入れ替える. 図にすると以下の図 3 ようになる.

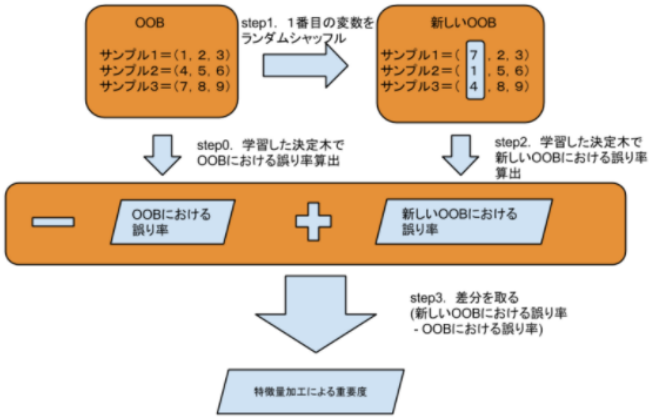

図 3 重要度の計算方法

図 3 のように、OOB における誤り率を計算し、どの程度精度が下がったかを指標とする. これを各決定木で行い、木あたりの平均を求める. この値が特徴量の重要度となる.

4 従来の手法と比較して

従来の手法の一つである SVM の欠点は学習が非常に非効率なことである. よって、データのサンプル数が多い場合、どのような問題でもオススメはできない. もっと言えば工業規模 (industry scale) で使うことは推奨できない. 研究室レベルや単純化された問題では他のアルゴリズムよりうまく動作する.

一方で、サンプルデータ数が十分に用意できる場合は、SVM ではなくランダムフォレストを使うほうが良い. 実際、Kinect の身体部位測定の判定にランダムフォレストベースのものが使われている.

3.1 メリット

- ノイズに強い
- 過学習を防ぐことができる
- 並列化が容易
- どの特徴量が重要かを知ることができる

3.2 デメリット

- パラメータが多い (木の数や使用する説明変数の数)
- 学習データ/説明変数をランダム抽出するので、データと変数が少なすぎるとうまく学習できない

5 まとめ

ランダムフォレストは集団学習の 1 つで、二分決定木を複数本作ることで学習器を構成していることが分かった. また、SVM と比較すると、より多くの学習データが用意できる場合は、ランダムフォレストに劣らない結果をえることができることが分かった. また、どの特徴量が重要であるかを知ることが出来る点で、ランダムフォレストは評価結果の考察が行いやすく、結果の改善が期待できると考えられる.

6 おわりに

今回調べてみて、多くの学習手法がある中でどの手法がどのようなデータ処理に向いているかは、臨機応変に見極めなければならないことが分かった. また、見極めたり議論を行う上で、それぞれの手法の特徴を理解しておくことが重要だ.

参考文献

[1] 1 ランダムフォレストの理論と重要な特徴量の選定
http://drilldripper.hatenablog.com/entry/2016/10/04/211245

[2] 2 機械学習ハッカソン
https://www.slideshare.net/

[3] 3 fMRI を用いた脳情報デコーディングに適した機械学習

[4] 4 Random Forest で計算できる特徴量の重要度
http://alfredplpl.hatenablog.com