ランダムフォレスト

1. ランダムフォレスト

1.2. 論文の概要

セクション 2 では, ランダムフォレストの理論的背景をいくつか示します. 大数の強法則を用いることで, 常に収束することが示されるので, オーバーフィッティングは問題ではない. ランダムフォレストの精度が個々の分類木の強度に依存することとその依存の程度を示すために, 1997 年のアミットとジェマンの分析を簡略化し, 拡張します (詳細はセクション 2).

セクション 3 では, 分割を決定するために, 各ノードで特徴量のランダムな選択を用いたフォレストを紹介します. 重要な問題は, どれだけの特徴量を各ノードで選択するかということです. 例として, 汎化誤差, 分類木の強度, 依存度を計算する. これらは out-of-bag とよばれ, それらの評価及び詳細はセクション 4 で行います. セクション 5 と 6 では 2 つの異なるランダムな特徴量の実験結果を示します. 1 つは元データからランダムに選択したもの, もう 1 つは元データをランダムに線形結合したものです. その結果は Adaboost と比較して優れています.

各ノードの特徴量の選択の数は結果に影響しないことが分かります. 通常, 1~2 の特徴量で最適な結果になります. この詳細やその強度と相関関係については, セクション 7 で実験します.

Adaboost はランダムな要素をもたず, 過去のアンサンブル学習の重み付けから, 再度トレーニングセットの前向き重み付けをすることで作ります.

しかし, 決定された乱数生成器が, 乱雑な良いコピーを生むので, Adaboost では後の段階でランダムフォレストを模倣しているというのが私の考えです. この考えの論拠は, セクション 8 で与えます.

診断や文書検索といった, 最近の重要な課題では, 多くの場合, 何百または何千といった入力 (それぞれの入力の情報量はは小さい) が与えられる. 単一の分類木でクラスのランダムな選択だと精度はほんのわずかな改善しか得られないでしょう. しかし, ランダムな特徴量を用いて作った木を組み合わせることで, 精度の改善をすることができます. セクション 9 では 1,000 個の変数入力と, 1,000 組の学習データ, 4,000 個のテストデータで実験を行います. ベイズ誤り率と同等の精度が達します.

多くのアプリケーションでは, ブラックボックスなランダムフォレストを理解することが必要です. セクション 10 では, その解明に, 変数の重要さの内部評価を計算して, それらを再利用することで着手する.

セクション 11 では, 回帰におけるランダムフォレストを考察する.　平均 2 乗汎化誤差の上界を導出します. 上界は, それぞれの木と比べたときの誤差の減少は, 残差とそれぞれの木の平均二乗誤差の相関関係に依存するということを示しています. 回帰の実験結果は, セクション 12 にあります. 結論は, セクション 13 にあります.

2. ランダムフォレストの精度の特徴

$h_1(\mathbf{x}), h_2(\mathbf{x}), ..., h_{\boldsymbol{K}}(\mathbf{x})$ を分類木とする. トレーニングセットをランダムなベクトル$\mathbf{Y}, \mathbf{X}$からランダムに抽出したものとし, マージン関数を次のように定義します.

$$mg(\mathbf{Y}, \mathbf{X}) = av_k I(h_k(\mathbf{X}) = Y) - \max_{j \neq Y} av_k I(h_k \mathbf{X}) = j).$$

ここで, $I(\cdot)$は指示関数である. このマージンは X, Y での平均投票数が, 他のクラスの平均投票数をどの程度上回っているかを表します. マージンが大きいほど分類に対する信頼度が高くなります. 汎化誤差は次式で与えられます.

$$PE^* = P_{\mathbf{X},Y}(mg(\mathbf{X}, Y) < 0)$$

$\mathbf{X}, Y$ の添字は, 確率 $\mathbf{X}, Y$ 上にあることを意味しています.

ランダムフォレストでは, $h_k(\mathbf{X}) = h(\mathbf{X}, \Theta_k)$ となります. 木が多くある場合, 汎化誤差は大数の強法則より以下のようになります.