Random Forests の寄与率を用いた効率的な特徴選択法の提案

EP09061 嶋崎 克也　　指導教授：藤吉弘亘

1. はじめに

識別器の構築には，高次元の特徴集合が用いられることが多い．高次元の特徴集合には，識別能力に貢献しない不必要な特徴次元が存在することがある．不必要な特徴次元は学習速度や学習モデルの可読性を低下させる原因となるため，次元数を削減する手法が提案されているが，特徴選択に時間を要するという問題がある．そこで，本研究では Random Forests[1] における寄与率を算出し，寄与率を基に特徴次元の削減を効率的に行うことを目的とする．

2. 提案手法

本研究では，Random Forests における特徴次元の寄与率を定義し，寄与率を用いて効率的な特徴選択を行う．

2.1. 寄与率の算出

Random Forests の学習過程で多くの学習サンプルを分割する分岐ノードで選ばれた特徴次元ほど識別能力への寄与が高いと考え，寄与率を定義する．寄与率 C は分岐ノード j で選ばれた i 番目の特徴次元 v_i がサンプルを分割した数の総数と全ての特徴次元が分割したサンプル数の割合として，次式のように定義する．

$$\text{寄与率 } C(v_i) = \sum_{t=1}^{T} \frac{\sum_{j \in f(v_i)} S_{t,j}}{\sum_{j \in J} S_{t,j}} \times 100 \qquad (1)$$

式 (1) の分子は，i 番目の特徴次元 v_i が分岐ノード j でサンプルを分割した数を算出する．構築した木の数 T 全て加算し，分割したサンプルの総数を求め，全ての分岐ノードに到達したサンプルの総数を用いて除算を行い算出する．図 1 に寄与率の算出方法を示す．

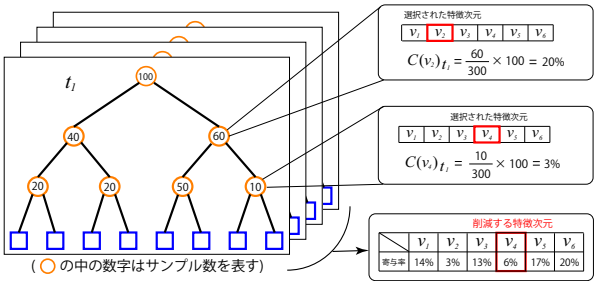

図 1：寄与率の算出

2.2. 寄与率を用いた特徴選択

寄与率を用いた特徴選択法として，逐次算出型とバッチ型の以下の 2 つのアルゴリズムを提案する．

・逐次算出型による特徴選択 (図 2(a))

逐次算出型は，Random Forests の決定木構築時に毎回寄与率を算出し，最も低い寄与率の特徴次元を削減する．

Step1 現在の特徴次元集合で Random Forests を構築
Step2 式 (1) より各特徴次元 v_i の寄与率 $C(v_i)$ を算出
Step3 構築した識別器の評価を行い，誤識別率を算出
Step4 特徴選択の終了条件を満たしていれば特徴選択を終了する終了条件を満たしていない場合は，最も寄与率の低い特徴次元の削減し，Step1 へ

・バッチ型による特徴選択 (図 2(b))

バッチ型は，全ての特徴次元を使った Random Forests の決定木構築の際に一回のみ寄与率を算出し降順リストを作成する．これを基に寄与率の低い特徴次元から削減する．

Step1 全ての特徴次元を用いて Random Forests を構築．式 (1) より寄与率 $C(v_i)$ を一回のみ算出し，降順リストの作成．Step3 へ
Step2 現在の特徴次元集合を用いて Random Forests を構築
Step3 構築した識別器の評価を行い，誤識別率を算出
Step4 特徴選択の終了条件を満たしていれば特徴選択を終了する．終了条件を満たしていない場合は，Step1 で作成した降順リストを基に最も寄与率の低い特徴次元の削減し，Step2 へ

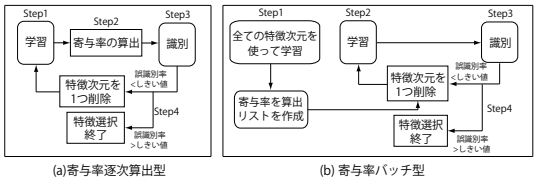

図 2：提案する特徴選択法

3. 評価実験

提案手法の有効性を示すため，従来法の SBS(Sequential Backward Selection)[2] 法と比較実験を行う．SBS 法は特徴次元を一次元削除して識別器を構築し，識別器の性能から削除した特徴次元を評価する手法である．

3.1. 実験概要

評価実験では，全ての特徴次元を用いた場合の誤識別率から 10%を超えた時点を特徴選択の終了条件として，誤識別率，特徴選択時間を比較する．評価実験には UCI Machine Learning Repository から 4 つのデータセットを用いる．

3.2. 実験結果

提案手法と SBS 法による誤識別率の推移と，提案手法と SBS 法が選択した特徴の共通する割合を図 3 に示す．誤識別率の推移の結果より，提案手法は従来法とほぼ同等の識別率で特徴次元を削減することができた．また，共通性においては，削減するごとに共通する割合は低下するが，80%削減した時点でも約半数が従来法と同じ特徴次元を選んでいる結果となった．提案する二手法では，逐次算出型がより多くの特徴次元を削減できた．逐次算出型は，毎回の Random Forests 構築の際に寄与率を算出し直す．これにより，特徴次元集合の変動を考慮した評価ができるため，識別精度に影響の無い特徴次元をバッチ型より多く削減できたと考えられる．

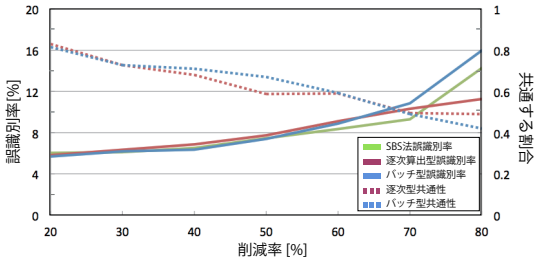

図 3：誤識別率の推移と共通性

次に，各データセットの特徴次元 1 つを削減するための処理時間を図 4 に示す．括弧内は削減個数を表す．提案手法は従来手法の SBS 法と比較して，特徴選択時間を大幅に短縮することができた．

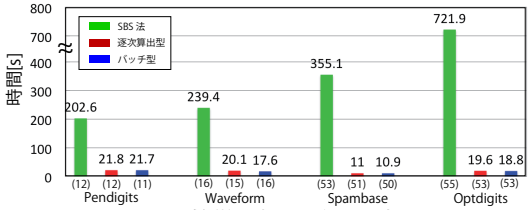

図 4：特徴選択に要する時間

4. おわりに

Random Forests の分岐ノードで選択された特徴次元の評価に寄与率を用いることで，識別精度に影響の無い特徴次元を効率的に削減することができた．今後は，特徴次元間の関係性を考慮した寄与率の算出法について検討し，より精度の高い特徴選択を目指す．

参考文献

[1] L. Breiman,“ Random Forests”, Machine Learning, vol. 45, pp.5-32, 2001.

[2] Marill, T, D. M. Green,“ On the effectiveness of receptors in recognition system ”, IEEE Trans. Inform. Theory 9, pp. 11-17, 1963.