社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

分岐ノードにおけるクラス間の分布を考慮した Random Forestsの高精度化

三品　陽平† 　山内　悠嗣† 　藤吉 弘亘††

† 中部大学　〒 487–8501 愛知県春日井市松本町 1200
E-mail: †{mishi,yuu}@vision.cs.chubu.ac.jp, ††hf@cs.chubu.ac.jp

あらまし　近年，パターン認識の分野において，大規模なデータベースから効率良く識別器を構築できる Random Forests が多く利用されている．Random Forests では，決定木の分岐関数の選択において分岐したサンプルの情報利得により評価している．しかし，情報利得はクラスの生起確率に基づいて算出されるため，分岐関数のしきい値とサンプルの分布の関係性が考慮されていない．そのため，分岐関数のしきい値とクラスが近く分布している場合，未知入力サンプルとしきい値の関係が反転し誤識別する可能性がある．そこで，本研究ではサンプルの分布に着目し，クラス間の距離を評価するために分離度を導入する．これにより，しきい値とクラス間が離れて分布するような汎化性能の高い分岐関数を選択できるため識別性能の向上が期待できる．

キーワード　Random Forests，分離度，マルチクラス識別器

Improving Random Forests by Introducing of Distribution Between Classes at Split Node

Yohei MISHINA†, Yuji YAMAUCHI†, and Hironobu FUJIYOSHI††

† Chubu University　Matsumoto-cho 1200, Kasugai-shi, Aichi, 487–8501 Japan
E-mail: †{mishi,yuu}@vision.cs.chubu.ac.jp, ††hf@cs.chubu.ac.jp

1. はじめに

近年，パターン認識の分野では，多数の決定木によって構成された Random Forests [1] が，画像分類やキーポイントマッチングに利用されている．Random Forests は，決定木の分岐関数を選択する際に情報利得を用いて，サンプルをどの程度分岐できるかを評価している．情報利得はクラスの生起確率に基づいて算出されるため，分岐関数のしきい値とサンプルの分布の関係性が考慮されていない．そこで，本研究では分岐したサンプル間の距離を考慮した分岐関数の選択法として分離度を導入し，識別器の汎化性能の向上を目的とする．

2. 提案手法

Random Forests [1] では，情報利得 ΔE が最大となる分岐関数を選択する．情報利得 ΔE は式 (1) により求める．

$$\Delta E = -\frac{I_l}{I_n}E(I_l) - \frac{I_r}{I_n}E(I_r) \tag{1}$$

情報利得は生起確率に基づいて算出されるため，左右に分岐したサンプルの分布は考慮されない．より汎化性能の高い識別器を学習するためには，分岐したサンプル間の距離を大きくする必要がある．そこで，本研究では分岐したサンプルの分布に着目する．サンプルの分布の関係を考慮するために，大津の 2 値化法 [2] の分離度を導入する．

2.1 分離度の導入

大津の 2 値化法の分離度は，2 クラス問題を対象としている．一方，Random Forests はマルチクラス問題も扱うため，提案手法ではマルチクラスに対応した分離度を導入する．図 1 にマルチクラスに対応した分離度のアイディアを示す．図 1 に示すように，しきい値により左右に分岐したサンプルを 2 クラスと見立てることで，マルチクラス問題において分離度を計ることができる．大津の 2 値化法ではサンプルごとに左右 2 クラスに分離するが，提案手法ではクラスごとに分離する．

2.2 クラス間分散と全分散の導出

分離度は，クラス内分散とクラス間分散の比である．しかし，クラス内分散の算出に要する計算量が多いため，しきい値に依存しない全分散を用いることで計算コストを低減する．以下に，クラス間分散と全分散の導出を示す．まず，ランダムに選択された特徴次元において各クラスの平均値 μ_i を求める．そして，ランダムに選択されたしきい値 Th とクラスの平均値 μ_i を用いて，式 (2) により擬似的に左右 2 クラスに分離する．

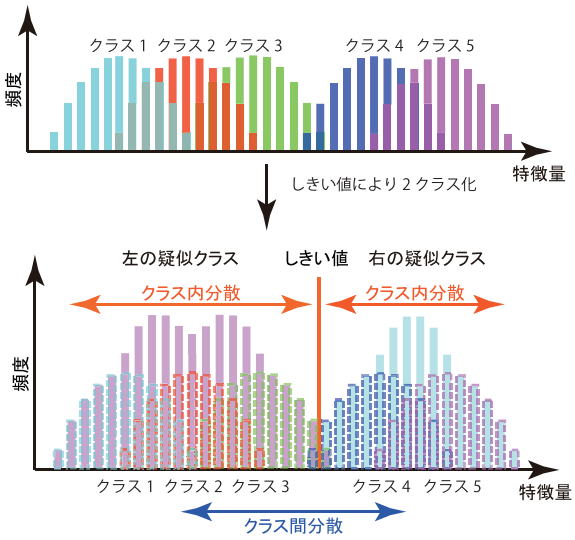


図 1　マルチクラスに対応した分離度のアイディア

$$\begin{cases} c_i \in \mathcal{C}_l & \text{if } \mu_i \leqq Th \\ c_i \in \mathcal{C}_r & \text{otherwise} \end{cases} \tag{2}$$

左右の疑似クラス $\mathcal{C}_l$，$\mathcal{C}_r$ の平均値 μ_l, μ_r を算出する．疑似クラスの平均値 μ_l, μ_r と全サンプルの平均値 μ_A を用いて，クラス間分散 σ_B^2 を式 (3) により求める．

$$\sigma_B^2 = \frac{n_l}{n}(\mu_l - \mu_A)^2 + \frac{n_r}{n}(\mu_r - \mu_A)^2 \tag{3}$$

ここで，n は分岐ノードのサンプル数である．そして，μ_A を用いて全分散 σ_A^2 を式 (4) により求める．

$$\sigma_A^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu_A)^2 \tag{4}$$

式 (3)，(4) より分離度は，$\eta = \sigma_B^2/\sigma_A^2$ と表される．全分散に対しクラス間分散が大きい場合に，汎化性能が最大になると考えられる．そのため，分離度が最大となる候補を分岐関数として選択する．識別時は，従来の Random Forests と同様に，各決定木の出力を平均し，事後確率が最大となるクラスに判別する．

3. 評 価 実 験

提案手法の識別性能を学習サンプル数を変化させて傾向を比較する．

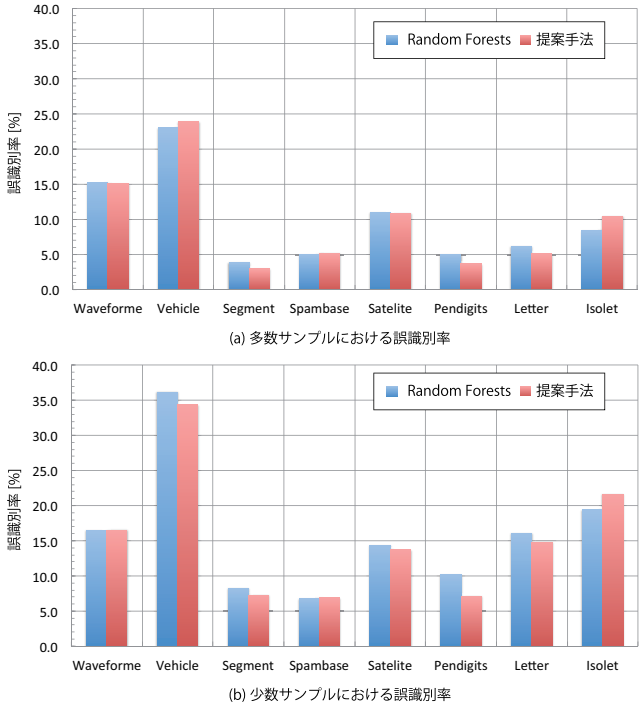
3.1 実 験 概 要

評価実験には，UCI Machine Learning Repository [3] の 8 つのデータセットを用いて，多数の学習サンプルと少数の学習サンプルにおける識別性能を比較する．少数の学習サンプルは，多数の学習サンプルから 20%ランダムに選択する．実験に用いた学習パラメータは，決定木の本数は 100 本，特徴量選択回数は特徴次元の平方根，しきい値選択回数は 10 回，サブセットのサンプル数は学習サンプル数の 66%とする．各学習パラメータにおいて，20 回試行した平均誤識別率を算出する．

3.2 実験結果と考察

従来法と提案手法のデータセットごとの誤識別率を図 2 に示す．多数の学習サンプルにおける識別性能では，従来法と提案手法はほぼ同程度である．少数の学習サンプルにおける識別性能では，提案手法が平均約 0.7%向上した．

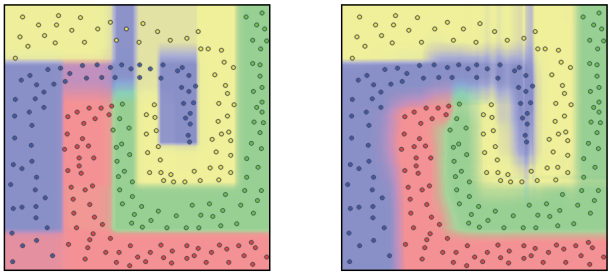
図 3 に人工データにおける識別境界の可視化例を示す．図 3(a) に示す情報利得を用いた分岐関数の選択は，分岐サンプルの分布を考慮できず直線的な識別境界である．また，識別境界付近にサンプルが分布しており，未知入力サンプルに対して汎化性能が低いと考えられる．一方，図 3(b) に示す分離度を用いた分岐関数の選択は，サンプルの分布に柔軟な境界である．分離度を用いて分岐関数を選択することにより，サンプルと識別境界との距離が大きくなり，汎化性能の高い識別器が構築することができた．



(a) 多数サンプルにおける誤識別率

(b) 少数サンプルにおける誤識別率

図 2　従来法と提案手法の誤識別率



(a) 情報利得を用いた分岐関数選択　(b) 分離度を用いた分岐関数選択

図 3　識別境界の可視化例

4. お わ り に

本稿では，Random Forests の分岐関数の選択に分離度を導入することを提案した．評価実験の結果，学習サンプルが少数の場合に従来法よりも汎化性能が向上する傾向があることを確認した．今後は，提案手法を画像分類やキーポイント分類等へ適用する予定である。

文　　献

[1] L. Breiman, “ Random Forests”, Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[2] 大津展之, “ 判別および最小 2 乗規準に基づく自動しきい値選定法 ”, 電子情報通信学会論文誌, vol. 63-D, no. 4, pp. 349-356, 1980.
[3] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/.