# 分離度を用いた分岐関数選択による Random Forests の高精度化

三品 陽平† 山内 悠嗣† 藤吉 弘亘‡
†‡ 中部大学
† {mishi, yuu}@vision.cs.chubu.ac.jp, ‡ hf@cs.chubu.ac.jp

**あらまし** 近年，機械学習やパターン認識の分野において，大規模なデータベースを用いて効率よく学習できる Random Forests が多く利用されている．Ranfom Forests は，決定木の分岐関数選択に分岐したサンプルの情報利得を用いて評価している．しかし，情報利得はクラスの生起確率に基づいているため，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．よって，分岐関数のしきい値とクラスの分布が近い場合，未知入力サンプル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，クラス間の分布の広さを評価するために分離度を導入することを提案する．これにより，分岐関数とクラス間の分布が広く汎化性能の高い分岐関数を選択することで識別性能の向上が期待できる．比較実験の結果，従来の Random Forests より約 3.39% の識別性能向上を確認した．

## 1 はじめに

近年，機械学習やパターン認識の分野では計算機の高性能化により，大規模なデータベースを容易に扱うことができるようになった．そこで，大規模なデータベースを用いたデータ解析や識別器の構築のニーズが高まっている [1]．識別器は大規模なデータベースを効率よく学習する必要があり，多数のカテゴリーを分類する高い識別性能も必要とされている．2 クラス識別器である Suport Vector Machine[2] や AdaBoost[3] を One-versus-Rest 法によりマルチクラスに識別することも可能である．また，AdaBoost の弱識別器をマルチクラス化した Multi-Class AdaBoost[4] も提案されている．しかし，これらの識別器は逐次型学習のため，大規模なデータベースを用いた学習に莫大な時間を要するという問題がある．この問題に対して，2001 年に L. Breiman らにより，大規模なデータベースを効率よく学習でき，識別性能の高いマルチクラス識別器として，Random Forests[5] が提案された．Random Forests の特徴は，複数の決定木で構成されたアンサンブル識別器である．これらの決定木は，学習サンプルから複数のサブセットを作成し，サブセットごとに決定木を学習する．独立した決定木を作成するため，各決定木を並列に学習することも可能である．そのため，画像認識の分野でも Random Forests を応用した手法が提案されている [6], [7].

Random Forests は，分岐関数選択に分岐したサンプルの情報利得を用いて評価している．しかし，情報利得では分岐関数のしきい値とサンプルの分布の関係性は考慮されない．サンプルがしきい値付近に分布するような分岐関数では，識別時に未知入力サンプルが反転して誤識別を起こす可能性が高い．

そこで本研究では，Random Forests の高精度化を目的とし，分岐関数選択に分離度を導入する．分離度は，分岐した左右のサンプル分布の広さを評価できる．これにより，従来法より汎化性能の高い分岐関数に選択することで，識別器全体の識別性能の向上が期待できる．

## 2 Random Forests

Random Forests[5] は学習サンプルを用いて複数の決定木を作成するアンサンブル識別器である．Random Forests は，大規模なデータベースからマルチクラス識別器を容易に作成できる．また，学習サンプルにブートストラップ法を適用したサブセットを用いて，各決定木を学習することで過学習を抑制している．しかし，学習サンプルからサブセットを作成するため，大量に学習サンプルが必要である．

各決定木はサンプルを分割する分岐ノードと確率密度分布を持つ末端ノードにより構成されている．分岐ノードの分岐関数の設計は基本的に自由である．分岐関数の選択は，複数候補の中から情報利得がもっとも高い候補に決定する．情報利得 $\Delta E$ は式 (1) より求める．

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

情報利得の算出には，主にエントロピーを用いる．エントロピー $E$ は式 (2) より求める．

$$E(I) = -\sum_{i=1}^{C} \mathrm{P}(c_i) \log \mathrm{P}(c_i) \tag{2}$$

ここで，$\mathrm{P}(c_i)$ は，クラス $c_i$ の生起確率を示す．
識別時は，未知入力サンプル $\boldsymbol{v}$ を各決定木に入力し，たどり着いた末端ノードに保存されている事後確率 $\mathrm{P}(c|l)$ を出力する．各決定木から出力された事後確率 $(P_1(c|l), P_2(c|l), \ldots, P(c|l))$ を用いて式 (3) により識別器全体の事後確率 $\mathrm{P}(c|\boldsymbol{v})$ を算出する．

$$\mathrm{P}(c|\boldsymbol{v}) = \frac{1}{T}\sum_{t=1}^{T}\mathrm{P}_t(c|l) \tag{3}$$

未知入力サンプルのクラス判別は式 (4) により求める．

$$C_i^* = \underset{c_i}{\operatorname{argmax}}\,\mathrm{P}(c_i|\boldsymbol{v}) \tag{4}$$

## 3　提案手法

Random Forests は分岐関数の選択に，情報利得を用いて分岐関数を評価している．情報利得は，左右に分岐したサンプルのエントロピーを用いて算出する．エントロピーは，各クラスの生起確率に基づいているため，情報利得を用いた評価ではサンプルが左右にいくつ分離するかの評価となる．しかし，未知入力サンプルに対する汎化性能を考慮すると，分岐関数のしきい値と学習サンプルが離れて分布している方が汎化性能が高い．汎化性能を高めるために，我々は分岐関数の選択に，大津の 2 値化法の分離度を導入する．大津の 2 値化法の分離度は，2 クラス問題の場合のみ適用できる．Random Forests に適用するためにマルチクラスに展開する．

### 3.1　大津の 2 値化法

大津の 2 値化法 [9] について簡潔に説明する．大津の 2 値化法は判別分析法とも呼ばれ，グレースケール画像を 2 値化する際に，最適なしきい値を自動的に選定する方法である．入力されるグレースケール画像のヒストグラムは，前景と背景の双峰性であると考えられる．大津の 2 値化法では，前景と背景のクラス内分散とクラス間分散の比を分離度とし，ヒストグラムの双峰を最大限分離するしきい値を自動で選定する．

### 3.2　マルチクラス問題への適応

大津の 2 値化法は，前景か背景の 2 クラス問題である．Random Forests では 2 クラス以上を扱うため，現状ではそのまま適用することはできない．そこで，サンプルの分離度の評価において，疑似的に左右に分岐したサンプルを 2 クラスと見なすことでマルチクラス問題へ展開する．
分岐関数は，しきい値を用いてサンプルを左右に分岐するため 2 値化と言える．そこで，マルチクラスを左右に分岐した 2 クラスと見立てることで，左右に分岐されたサンプルの分離度を計量できる．図 1 にアイディアの模式図を示す．左右のクラスに分類する基準は，各クラスの特徴量の平均値としきい値の大小関係で分類する．この左右に分類する基準が，大津の 2 値化法と大きく異なる点である．大津の 2 値化法はサンプルごとに分類する．一方，提案手法ではクラスごとに分離する．

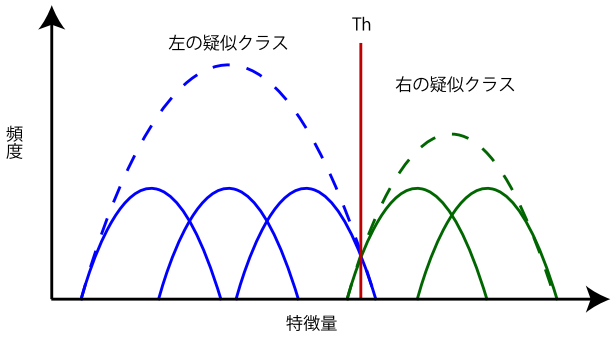



図 1: アイディアの模式図

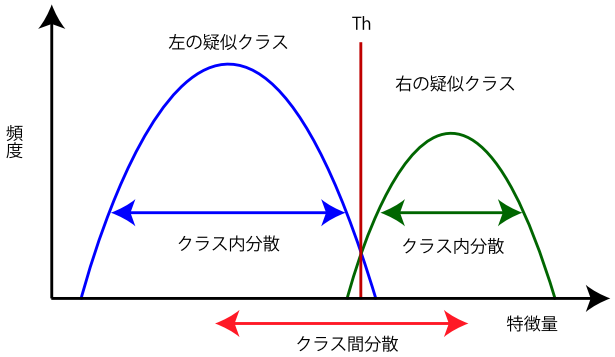



図 2: クラス内分散とクラス間分散の例

#### 3.2.1　クラス間分散とクラス内分散の導出

図 2 にクラス内分散とクラス間分散の例を示す．クラス内分散は，左右それぞれの疑似クラスの広がりである．クラス間分散は，左右のクラスの頻度分布がどの程度広がっているかを示す．
以下に，クラス内分散とクラス間分散の導出を示す．ある分岐ノードにおいて，クラス数 $C$ のサンプル集合 $\mathcal{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n)$ があるとする．まず，ランダムに選択された特徴量 $x$ において，各クラスの平均値を求める．そして，ランダムに選択されたしきい値 $Th$ が与えられ，式 (5) により擬似的に左右 2 クラスに分類する．

$$\begin{cases} c_i \in \mathcal{C}_l & \text{if } \ \mu_i \leq Th \\ c_i \in \mathcal{C}_r & \text{otherwise} \end{cases} \tag{5}$$

ここで，$\mu_i$ はクラス $c_i$ の平均値を示す．$\mathcal{C}_l$ は左に分岐されたクラスの集合，$\mathcal{C}_r$ は右に分岐したクラスの集合

である．疑似クラス $\mathcal{C}_l$，$\mathcal{C}_r$ の平均値 $\mu_l, \mu_r$ は式 (6), (7) となる．

$$\mu_l = \frac{1}{n_l} \sum_{\boldsymbol{s}_j \in \mathcal{C}_l} x_j \tag{6}$$

$$\mu_r = \frac{1}{n_r} \sum_{\boldsymbol{s}_j \in \mathcal{C}_r} x_j \tag{7}$$

ここで，$n_l, n_r$ は，それぞれに分類されたサンプル数である．そして，左右の疑似クラスの分散 $\sigma_l^2, \sigma_r^2$ は式 (8), (9) により表される．

$$\sigma_l^2 = \frac{1}{n_l} \sum_{\boldsymbol{s}_j \in \mathcal{C}_l} (x_j - \mu_l)^2 \tag{8}$$

$$\sigma_r^2 = \frac{1}{n_r} \sum_{\boldsymbol{s}_j \in \mathcal{C}_r} (x_j - \mu_r)^2 \tag{9}$$

これらの式より，疑似クラスのクラス内分散 $\sigma_W^2$ は式 (10) となる．

$$\sigma_W^2 = \frac{n_l}{n}\sigma_l^2 + \frac{n_r}{n}\sigma_r^2 \tag{10}$$

また，疑似クラスのクラス間分散 $\sigma_B^2$ は式 (11) となる．

$$\sigma_B^2 = \frac{n_l}{n}(\mu_l - \mu_A)^2 + \frac{n_r}{n}(\mu_r - \mu_A)^2 \tag{11}$$

また，ノード全体のサンプル数を $n$ とすると，全体の平均値 $\mu_A$ は式 (12) となる．

$$\mu_A = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{12}$$

平均値により全分散 $\sigma_A^2$ は式 (13) と表される．

$$\sigma_A^2 = \frac{1}{n}\sum_{j=1}^{n} (x_j - \mu_A)^2 \tag{13}$$

#### 3.2.2 分離度の定義

分岐関数を評価するために，左右に分岐したサンプルの分離度を算出する．左右のクラスではクラス内分散が小さく，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．つまり，サンプルの分離度は式 (14) から求められる．

$$\lambda = \frac{\sigma_B^2}{\sigma_W^2} \tag{14}$$

しかし，クラス内分散は計算量が多いため，分離度に，しきい値により変化しない全分散を用いる．式 (15) より分離度は，クラス間分散と全分散の比とする．

$$\eta = \frac{\sigma_B^2}{\sigma_A^2} \tag{15}$$

式 (15) を用いてクラス間の分離度を計量し，分離度が最大となる分岐関数を選択する．

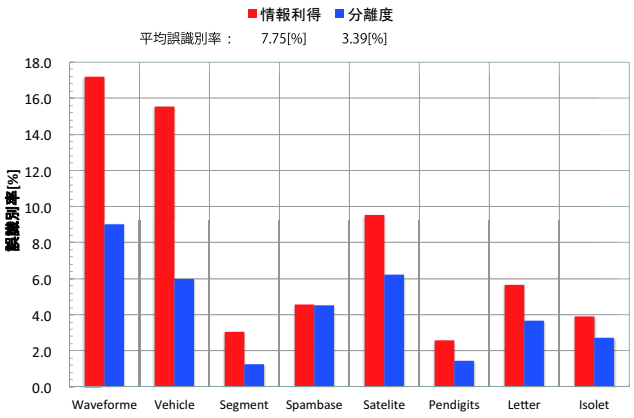



図 3: 従来法と提案手法の誤識別率

## 4 評価実験

従来手法と提案手法の識別性能を比較する．各データセットを用いて，10 回試行し平均誤識別率を算出する．

### 4.1 実験概要

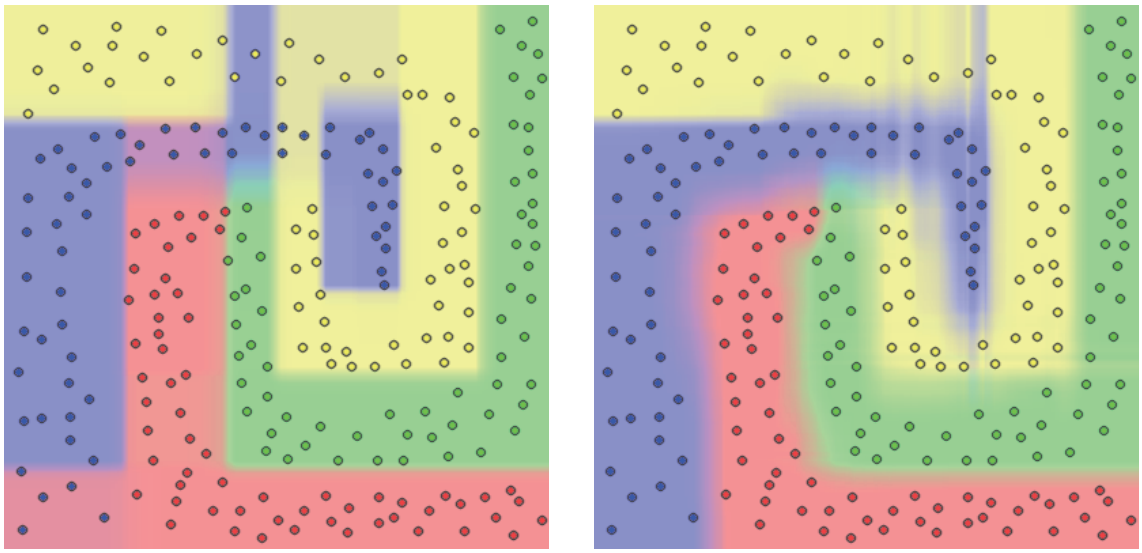
今回の実験に使用するデータセットを表 1 に示す．これらのデータセットは，機械学習アルゴリズムのベンチマークデータセットであり，UCI Machine Learning Repository[10] において公開されている．また，実験に用いた学習パラメータは，決定木の本数は 100 本，特徴量選択回数は特徴次元の平方根，しきい値選択回数は 10 回である．そして，サブセットのサンプル数は学習サンプル数の 25%とする．

### 4.2 実験結果と考察

従来法と提案手法の誤識別率を図 3 に示す．グラフの縦軸に誤識別率を，横軸に各データセットを示す．赤のビンが従来法を示し，青のビンが提案手法の誤識別率を示す．グラフの値が低いほど識別性能が高いことを示す．実験結果より，提案手法は Vehicle dataset のとき最大で約 9.57% 識別性能が向上した．また，平均で約 3.39% 識別性能が向上した．分岐関数選択に分離度を用いることで，識別性能が向上することを確認した．
図 4 に，識別境界の可視化画像を示す．図 4(a) に示す情報利得を用いた分岐関数選択は，サンプルの分布の広さを考慮できず直線的な境界である．また，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 4(b) に示す分離度を用いた分岐関数選択は，サンプルの分布に柔軟な分離境界を作成している．これらの効果により，識別性能が向上していると考える．

表 1: データセットの概要

| Dataset | 特量次元数 | クラス数 | 学習サンプル数 | テストサンプル数 |
|---|---|---|---|---|
| Waveforme | 21 | 3 | 300 | 4700 |
| Vehicle | 18 | 4 | 761 | 85 |
| Segment | 19 | 7 | 2079 | 231 |
| Spambase | 57 | 2 | 3221 | 1380 |
| Satelite | 36 | 6 | 4435 | 2000 |
| Pendigits | 16 | 10 | 7494 | 3498 |
| Letter | 16 | 26 | 10000 | 10000 |
| Isolet | 617 | 26 | 6238 | 1559 |



(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 4: 識別境界の可視化画像

## 5　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．情報利得を用いた分岐関数選択との比較実験の結果，識別性能の向上を確認した．今後は，末端ノードにおける確率密度関数の作成方法に着目し更なる識別性能の向上を目指す．

## 参考文献


[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: a large-scale hierarchical image database”, In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[2] C. Cortes, and V. Vapnik, “Support vector networks”, Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[3] Y. Freund, and R. E. Schapier,“ Experiments with a new boosting algorithm”, In Proceedings of the Thirteenth International Conference on Machine Learning, 1996.

[4] J. Zhu, and H. Zou, and S. Rosset, and T. Hastie,“ Multi-class adaboost”, Statistics and Its Interface, vol. 2, pp. 349-360, 2009.

[5] L. Breiman,“ Random Forests”, Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[6] J. Shotton, M. Johnson and R. Cipolla,“ Semantic texton forests for image categorization and segmentation”, Computer Vision and Pattern Recognition, 2008.

[7] M. Ozuysal, M. Calonder, V. Lepetit, P. Fua,“Fast Keypoint Recognition using Random Ferns”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 3, pp. 448-461, 2010.

[8] P. Geurts, and D. Ernst, and L. Wehenkel,“ Extremely randomized trees”, Machine Learning, vol. 63, no. 1, pp. 3-42, 2006.

[9] 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, no. 4, pp. 349-356, 1980.

[10] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/.