Vol.2014-CVIM-192 No.30
2014/5/15

Random Forest を用いた能動学習における有効なサンプル選択

村田 隆英[†1,a)]　三品 陽平[†1]　山内 悠嗣[†1]　山下 隆義[†1]　藤吉 弘亘[†1,b)]

概要：能動学習は，インタラクティブに新しいサンプルを選択してラベルを付与し，学習に用いることで，より良い識別境界を求めるアプローチである．ラベルを付与するサンプル選択の際に，Vote Entropy を用いた手法では類似したサンプルが選択されるため，学習の効率が悪いという問題がある．本研究では Random Forest を用いたサンプルの密度推定を行い，密度分布の類似度を考慮したサンプルの選択法を提案する．提案手法は，類似したサンプルの選択を抑制することで，少数のサンプルによる学習の効率化が期待できる．評価実験より，従来の能動学習におけるサンプル選択法と比較し，提案手法はより少ないサンプル数で高い性能を持つ識別器の構築が可能であることを確認した．

Effective Sample Selection Method Using Random Forest for Active Learning

Ryuei Murata[†1,a)]　Yohei Mishina[†1]　Yuji Yamauchi[†1]　Takayoshi Yamashita[†1]
Hironobu Fujiyoshi[†1,b)]

Abstract: Active learning is an approach to make a better discriminant boundary by interactively collecting new samples. Since conventional sample selection methods such as vote entropy collect similar samples for active learning, efficiency of the learning is not good. In this paper, we propose a method for selecting samples based on the similarity of density maps obtained by density forests. We confirmed that the proposed method makes better discriminant boundary with smaller number of samples used for the learning.

1. はじめに

画像認識の分野において，大規模な画像データベースを用いて機械学習により汎化性能の高い識別器を実現する必要性が高まっている [1]．画像データベースの肥大化に伴い，各サンプルに対するラベル付与のコストが増加している．ラベル付与のコストを削減する方法として，1 サンプルあたりのラベル付与のコストを削減する方法 [2] とサンプル数を削減する方法 [3] の 2 つのアプローチが提案されている．1 サンプルあたりのラベル付与のコストを削減する方法は，クラウドソーシングによる不特定多数の人に業務を委託するという新しい雇用形態を利用する．クラウドソーシングの一つに Amazon Mechanical Turk[2] がある．Amazon Mechanical Turk ではリクエスターが報酬を支払うことによって，インターネットを通じて不特定多数のワーカーにサンプルへのラベルを付与してもらい，データセットを作成する．しかし，作成するデータセットの専門性が高い場合，専門知識を有する人にしかラベルの付与を行うことができない．また，誤ったラベルを付与することがあり，ラベルの付与における信頼性の問題が存在する．

一方，サンプル数を削減する方法としては，識別境界の決定に寄与すると考えられるサンプルを選択し，ラベルを付与して学習に用いる．これを繰り返すことで，少数のラベル付きサンプルを用いてより良い識別性能を得ることが可能となる．このような学習法を能動学習 [3] と呼ぶ．能動学習では，ラベルを付与したサンプルは識別境界に大きな影響を与えるため，サンプルの選択が重要である．しか

†1 現在，中部大学
Presently with Chubu University
a) mryua@vision.cs.chubu.ac.jp
b) hf@cs.chubu.ac.jp

し，従来の能動学習では，サンプルにラベルを付与する際に，各サンプル毎の曖昧さのみを指標としているため，類似したサンプルが選択される場合があり，能動学習の効率が低下する恐れがある．

そこで，本研究では Random Forest[5] のフレームワークを用いた密度推定とラベル伝播 [6] を行うことで，能動学習において有効なサンプルを選択する．少数のラベル付きサンプルのみでは，正しく識別境界を決定することはできない．そのため，ラベル付きサンプルの情報をラベル無しサンプルに伝播し，より良い識別境界を決定することで少ないラベル付きサンプルでの学習効率を向上させ，効率良くサンプルを選択する．また，密度分布に着目し，Random Forest を用いた能動学習における有効なサンプル選択法を提案する．密度分布の類似度によるサンプル選択を用いることで，ラベル付きサンプルの追加回数の削減を実現する．

2. 能動学習における従来のサンプル選択法

教師あり学習において，性能が高い識別器を学習するためには，膨大なラベル付きサンプルが必要である．しかしながら，サンプルに対してラベルを付与するためには，時間や労力などのコストが高いという問題がある．能動学習 [3] では識別境界の決定に寄与しそうな少数のサンプルのみにラベルを付与することで，効率良く識別器を学習することを目的としている．この際，重要となるのは，どのサンプルにラベルを付与するかというサンプル選択である．能動学習のサンプルの選択法として，Uncertainty Sampling[7] や，Query-By-Committee[10] がある．

2.1 Uncertainty Sampling におけるサンプル選択

能動学習におけるサンプル選択法として，Uncertainty Sampling[7] がある．この手法は，曖昧な識別結果であるサンプルに対してラベルを付与することを目的としている．このようなサンプルにラベルを付与することで，識別境界付近に存在するようなサンプルを正しく識別可能な識別器を学習できる．Uncertainty Sampling における識別結果の曖昧さを測る Least Confident, Margin Sampling[8], Entropy[9] の 3 つの指標が提案されている．

・Least Confident

Least Confident では，識別結果のクラス確率が最も小さいサンプルを曖昧なサンプルとする手法である．式 (1) により，サンプル中の最も曖昧なサンプルを求めることができる．

$$x^{\star}_{LC} = \arg\max_{x}(1 - P_\theta(\hat{y}|x)) = \arg\min_{x} P_\theta(\hat{y}|x) \quad (1)$$

ここで，$\hat{y}$ はモデル θ の下で最も属するであろうクラスを表し，x は入力されたサンプルを表す．

・Margin Sampling

Margin Sampling では，サンプルの 1 番目に確率の高いクラス確率と 2 番目に確率の高いクラス確率の差を指標とし，差が最も小さいサンプルをラベル付与の対象とする手法である．式 (2) により，サンプル中の最も曖昧なサンプルを求めることができる．

$$x^{\star}_{M} = \arg\min_{x} \left(P_\theta(\hat{y_1}|x) - P_\theta(\hat{y_2}|x)\right) \quad (2)$$

ここで，$\hat{y_1}$ はサンプルが最も属するであろうクラスを，$\hat{y_2}$ はサンプルが二番目に属すであろうクラスを表している．

・Entropy

Entropy では，サンプルの全てのクラスに対するエントロピーを計算し，予測分布のエントロピーが最大のサンプルを，最も曖昧なサンプルとして選択する手法である．式 (3) により，サンプル中の最も曖昧なサンプルを求めることができる．

$$x^{\star}_{H} = \arg\max_{x} \left(-\sum_{i} P_\theta(\hat{y_i}|x) \log P_\theta(\hat{y_i}|x)\right) \quad (3)$$

ここで $\hat{y_i}$ はサンプルのすべてのクラスを表している．

2.2 Query-By-Committee におけるサンプル選択

アンサンブル学習を前提とするサンプル選択法として Query-By-Committee[10] がある．この手法は同じモデルで異なるパラメータからなる複数の識別器を学習し，それぞれの識別器からの投票 (識別結果) が分かれるサンプルにラベルを付与する．このようなサンプルは識別結果が明瞭でないサンプルであるため，このようなサンプルに対してラベルを付与して学習すると，識別境界の形成に大きく寄与する．投票がどれくらい割れたかを測る指標として Vote Entropy[11] が提案されている．

・Vote Entropy

Vote Entropy は，投票結果情報のエントロピーを算出し，エントロピーが最大のサンプルを選択する手法である．式 (4) により，曖昧なサンプルを選択する．

$$x^{\star}_{VE} = \arg\max_{x} \left(-\sum_{i} \frac{V(y_i)}{C} \log \frac{V(y_i)}{C}\right) \quad (4)$$

ここで C は識別器の数，$V(y_i)$ はラベル y_i を予測した識別器の数を表している．

2.3 従来のサンプル選択法の問題点

能動学習における従来のサンプル選択法では，ラベルを付与するサンプルの選択の際にサンプル間の分布を考慮しないため，**図 1** に示すようにラベルを追加しても，類似したサンプルが選択される場合があり，能動学習の効率が低下する恐れがある．そこで，サンプルの分布を考慮したサンプルの選択を行うために，Random Forest のフレームワークである密度推定を用いる．

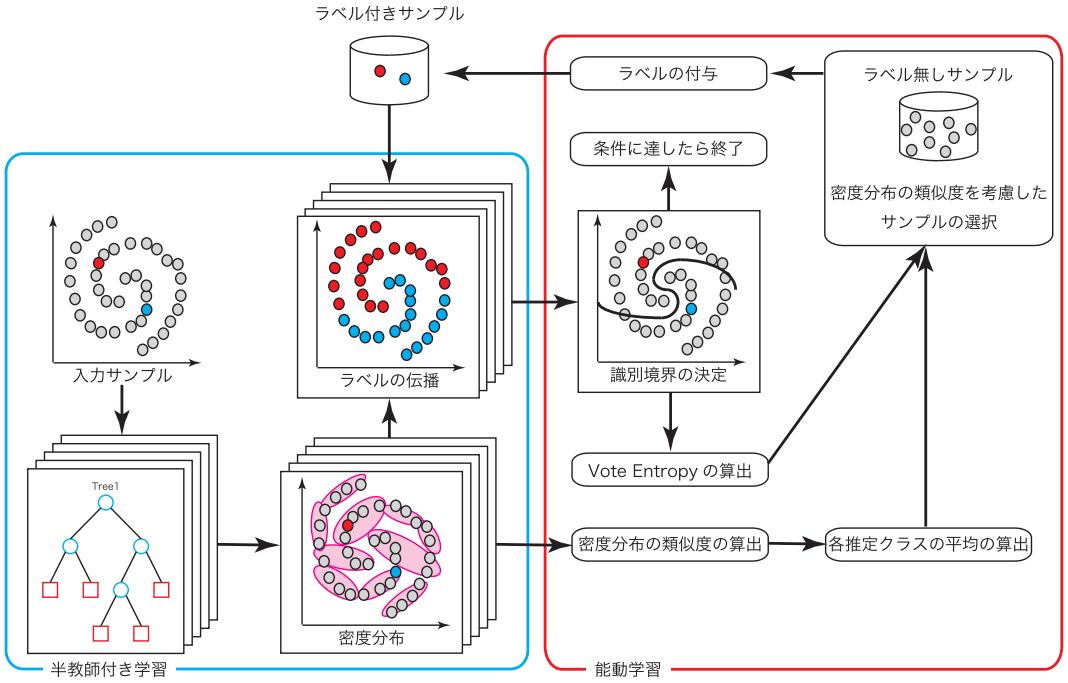


図 2 提案手法の流れ
Fig. 2 A flow of the proposed method.

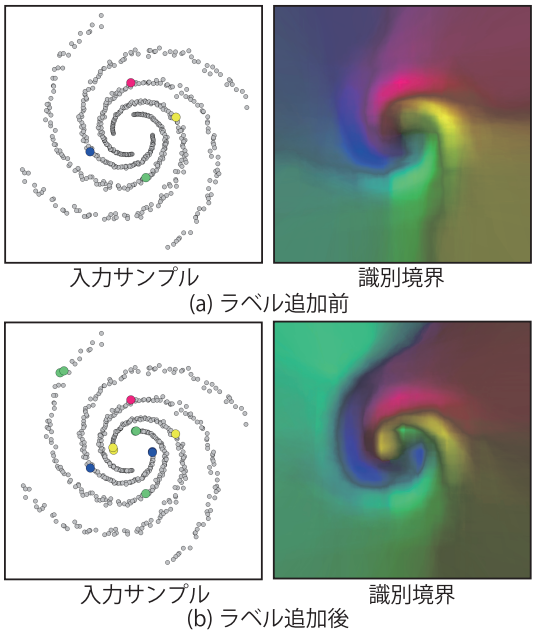


図 1 サンプル選択の問題
Fig. 1 Problem in sample slection

3. 提案手法

本研究では，Random Forest のフレームワークである密度推定 (Density Forest[6]) の結果に着目し，類似したサンプルの選択を抑制したサンプル選択法を提案する．**図 2** に提案手法の流れを示す．提案手法では，全サンプルを用いて Density Forest により密度推定を行う．密度推定の結果より，ラベル無しサンプルにラベルを伝播することで半教師付き学習を行う．また，密度推定の結果をサンプルの選択に用いることで，より効率良くサンプルを選択してラベルを付与する．

3.1 Density Forest による密度推定

サンプルの密度を推定するために，ラベル付きサンプル集合 $\mathcal{S}^{(s)}$ とラベル無しサンプル集合 $\mathcal{S}^{(u)}$ を合わせた全サンプル集合 $\mathcal{S}$ を用いて Density Forest を構築する．密度木の各ノードは，目的関数 I_j が最大となる特徴量と閾値でサンプル集合 $\mathcal{S}_j$ 分岐する．目的関数 I_j は式 (5) で定義される．

$$I_j = \log(|\Lambda(\mathcal{S}_j)|) - \sum_{\mathrm{i}\in\{\mathrm{L,R}\}} \frac{|\mathcal{S}_j^i|}{|\mathcal{S}_j|} \log(|\Lambda(\mathcal{S}_j^i)|) \qquad (5)$$

ここで，Λ は d 次元の特徴量の共分散行列，$\mathcal{S}_j$ はノード j に到達した学習サンプル集合を示し，$\mathcal{S}_j^{\mathrm{L}}, \mathcal{S}_j^{\mathrm{R}}$ は，それぞれ左または右に分岐したサンプル集合を示す．Density Forest ではラベル無しサンプルも扱うため，情報利得を算出する際にガウス分布のエントロピーを使用する．すなわち，分岐したサンプル集合の分散が小さくなるように分割される．これを末端ノードまで繰り返す．末端ノードの終了条件は，到達したサンプル数が一定の数より少ない場合もしくは，一定の深さに到達した場合とする．各末端ノードには，到達したサンプルの特徴量の各次元の平均値と特徴量の共分散行列を保存する．

また，各末端ノードに到達した学習サンプルを 1 つの多変量ガウス分布 $\mathcal{N}(\mathbf{v}; \mu_{l(\mathbf{v})}, \Lambda_{l(\mathbf{v})})$ で表す．t 番目の木の出力は式 (6) のように表される．

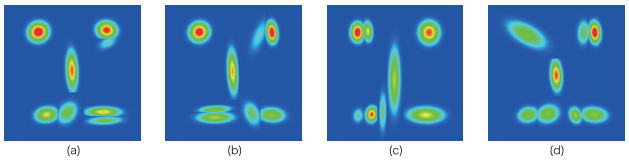


図 3 各木の断片的なガウス分布
Fig. 3 Density map for each tree

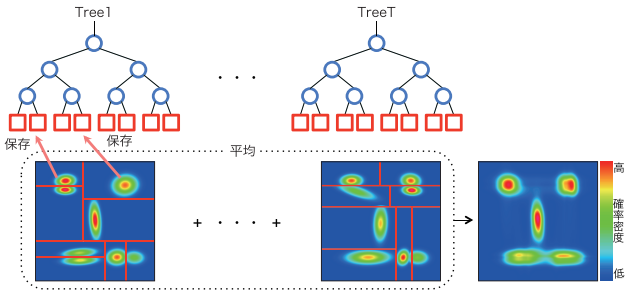


図 4 各木の出力の統合
Fig. 4 Averege of outputs

$$p_t(\mathbf{v}) = \frac{\pi_{l(\mathbf{v})}}{Z_t}\mathcal{N}(\mathbf{v}; \mu_{l(\mathbf{v})}, \Lambda_{l(\mathbf{v})}) \tag{6}$$

ここで，l は末端ノード，μ_l は末端ノードに到達したすべてのデータの平均値を示し，Λ_l は関連する共分散行列を示す．π_l は末端ノードに到達した学習サンプルと全体の学習サンプルの割合 $\pi_l = \frac{S_l}{S}$ を示す．**図 3** に各木の密度推定の結果を示す．このようにいくつかのガウス分布から構成される多変量ガウス分布で表される．また，各ガウス分布は各末端ノードの境界線上で切り捨てられているため，分布の重なりがある場合には分布の積分値が必ずしも 1 にならない．そこで，正規化をするために分配関数 Z_t を導入する．分配関数 Z_t を式 (7) に示す．

$$Z_t = \int_{\mathbf{v}} \pi_{l(\mathbf{v})}\mathcal{N}(\mathbf{v}; \mu_{l(\mathbf{v})}, \Lambda_{l(\mathbf{v})})d\mathbf{v} \tag{7}$$

また，式 (7) は式 (8) の数値積分によって近似できる．

$$Z_t \approx \Delta \cdot \sum_i \pi_{l(\mathbf{v})}\mathcal{N}(\mathbf{v}_i; \mu_{l(\mathbf{v}_i)}, \Lambda_{l(\mathbf{v}_i)}) \tag{8}$$

ここで，$\mathbf{v_i}$ は特徴空間上の特徴量，Δ は特徴量の分解能を示す．分解能が高いほど正確な近似を行うが，計算コストは高くなる．Density Forest による密度推定 $p(\mathbf{v})$ は，学習サンプルを密度木に入力し，複数の密度木の出力の平均値を式 (9) によって求める．

$$p(\mathbf{v}) = \frac{1}{T}\sum_t^T p_t(\mathbf{v}) \tag{9}$$

図 4 に示すように，各密度木の結果を統合することで Density Forest は複雑なデータ集合に対する密度推定を行うことができる．Density Forest による最終的な密度推定の結果を**図 5** に示す．

3.2 ラベル伝播

次に，各密度木の密度分布に沿うようにラベルを伝播す

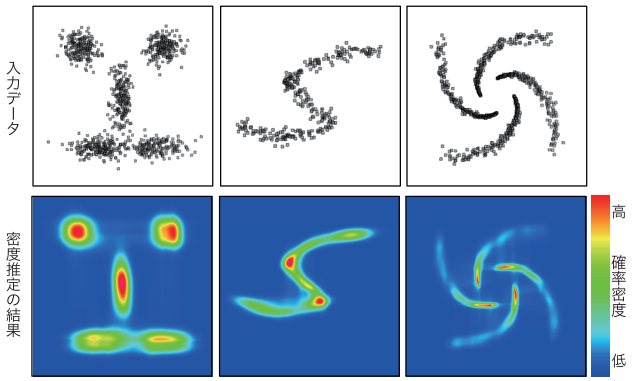


図 5 Density Forest による密度推定
Fig. 5 Examples of density estimation

ることで，ラベル無しサンプルにラベルを付与する．ラベルの伝播はラベル無しサンプルが無くなるまで繰り返す．ラベル付きサンプルからラベル無しサンプルへのラベルの伝播は以下の式を用いる．

$$c(\mathbf{v}^{\mathrm{u}}) \Leftarrow c\left(\arg\min_{\mathbf{v}^{\mathrm{l}} \in \mathcal{L}} D(\mathbf{v}^{\mathrm{u}}, \mathbf{v}^{\mathrm{l}})\right) \quad \forall \mathbf{v}^{\mathrm{u}} \in \mathcal{U} \tag{10}$$

ここで，関数 $c(\cdot)$ は入力のラベル無しサンプルに伝播するクラス，$\mathcal{U}$ はラベル無しサンプル集合を示す．また，サンプル間の距離計算には局所距離を用いる．しかし，局所距離ではラベル無しサンプルと追加されたサンプルとの距離になるため，局所距離に追加されたサンプルが今まで辿ってきた初期のラベル付きサンプルまでの距離 $l_{\mathbf{v}^{\mathrm{l}}}$ を加算し，測地線距離 $D(\cdot,\cdot)$ とする．これにより，分布に沿った距離を測ることができる．

$$D(\mathbf{v}^{\mathrm{u}}, \mathbf{v}^{\mathrm{l}}) = \min_{\mathbf{v}^{\mathrm{u}} \in \mathcal{U}} d(s_i, s_j) + l_{\mathbf{v}^{\mathrm{l}}} \tag{11}$$

次に，局所距離を計算する．今回は局所距離としてマハラノビス距離を用いる．ラベル付きサンプルとラベル無しサンプルがそれぞれ到達した末端ノードに保存された共分散行列を用いて，局所距離 $d(\cdot,\cdot)$ を算出する．局所距離は式 (12) のように定義される．

$$d(s_i, s_j) = \frac{1}{2}\left(\mathbf{d}_{ij}^{\mathrm{T}}\Lambda_{l(\mathbf{v}_i)}^{-1}\mathbf{d}_{ij} + \mathbf{d}_{ij}^{\mathrm{T}}\Lambda_{l(\mathbf{v}_j)}^{-1}\mathbf{d}_{ij}\right) \tag{12}$$

ここで，s_i, s_j をそれぞれ，ラベル無しサンプルの座標，ラベル付きサンプルの座標とし，$\mathbf{d}_{ij} = s_i - s_j$ はラベル無しサンプルとラベル付きサンプルの座標値の差，$\Lambda_{l(v_i)}^{-1}$ は逆行列，$\Lambda_{l(v_i)}$ はラベル無しサンプル $\mathbf{v}_u$ が到達した末端ノードの共分散行列を示し，$\Lambda_{l(v_j)}$ はラベル無しサンプル $\mathbf{v}_l$ が到達した末端ノードの共分散行列を示す．ラベルを伝播したラベル無しサンプルをラベル付きサンプルとして追加していくことで，初期のラベル付きサンプルからラベルの伝播が広がるように伝播する．伝播結果により末端ノードのクラス分布を作成することで，Random Forest と同様に識別を行うことができる．

3.3 曖昧さと密度分布の類似度によるサンプルの選択

サンプル選択に必要な曖昧さと密度分布の類似度を算出をする．**図 6** に各木の密度分布と各サンプルの密度分布の類似度を示す．学習結果から，ラベル無しサンプル集合 $\mathcal{S}^{(u)}$ に属している $\boldsymbol{x}_i$ が入力されたときの Vote Entropy の値 $VE(\boldsymbol{x}_i)$ と $\boldsymbol{x}_i$ が到達した各木の末端ノードが持つ密度分布の類似度 $D(\boldsymbol{x}_i)$ を算出する．複数の密度分布間の類似度 $D(\boldsymbol{x}_i)$ は，シャノンの情報量を用いた JS-Divergence により式 (13) により求める．

$$D(\mathcal{N}_1, \mathcal{N}_2, \cdots, \mathcal{N}_T) = H\left(\sum_{t=1}^{T} \mathcal{N}_t\right) - \sum_{t=1}^{T} H\left(\mathcal{N}_t\right) \quad (13)$$

ここで，$\mathcal{N}_t$ は t 本目の決定木におけるサンプル $\boldsymbol{x}_i$ の密度分布，$H(\cdot)$ はシャノンの情報量を示す．図 6(b) に各サンプルの密度分布の類似度を示す．Vote Entropy の値 $VE(\boldsymbol{x}_i)$ と，密度分布の類似度 $D(\boldsymbol{x}_i)$ を用いてサンプルを選択する．ここでは，$D(\boldsymbol{x}_i)$ から類似度が高いサンプル集合と低いサンプル集合に分け，それぞれ $VE(\boldsymbol{x}_i)$ の値が最大となるサンプルを選択する．密度分布の類似度を考慮したサンプルを選択することで，類似したサンプルの選択を抑制できる．

3.4 ラベルの再伝播によるクラス分布の更新

選択されたサンプルに対して人手によりラベルを付与する．本手法では，各推定クラスに対して 1 度に 2 個のサンプルが選択される．ラベルの選択の際に推定されたクラスを用いるため，追加されるサンプルのラベルには偏りが生じる場合もある．

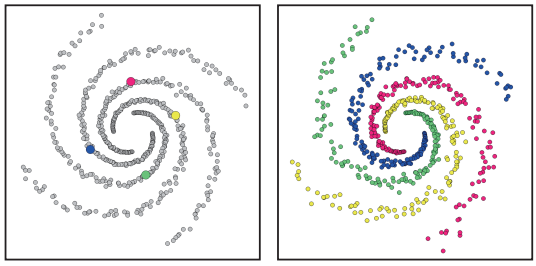
ラベルを付与したサンプルをラベル付き学習サンプルに追加し，ラベルの再伝播を行う．このとき，木の再構築は行わず各木の末端ノードのクラス分布のみ更新する．クラス分布を更新することで，識別結果が変化する．一定の条件に達するまでサンプルの選択からクラス分布の更新を繰り返すことで識別境界を更新する．

4. 評価実験

評価実験では，従来のサンプル選択法である Least Confident，Margin Sampling，Entropy，Vote Entropy の 4 手法と各手法に提案手法を加えた方法を比較する．

4.1 実験概要

従来法では，1 度に付与するラベルの追加数を各クラス 1 個の場合と 2 個の場合も比較する．比較を容易にするために，全手法において同条件となるようにラベル伝播の際に用いる密度推定の結果は同じものを用いる．RF のパラメータである木の数は 400 本，木の深さは 10，特徴次元の選択回数は 1，しきい値の選択回数は 10 回，末端ノードの最小サンプル数は 20 個とする．ラベルの追加の終了条件は識別率が一定の値に達した場合とする．また，評価実験を行うためのデータセットには Microsoft Research で公開されている，**図 7** に示す学習用と識別用の 2 次元のスパイラルデータ [6] を使用する．色付きの点がラベル付きサンプルを表し，灰色のサンプルがラベル無しサンプルを表す．



(a) 学習用サンプル　(b) 識別用サンプル

図 7 使用するスパイラルデータ
Fig. 7 Spiral data

4.2 実験結果

評価実験の各手法の結果を**図 8** に示す．提案手法は，従来法よりも少ないラベルの付与回数で目標である教師付き学習と同等の識別率を得ることができた．特に，Entropy と Vote Entropy において，ラベルの追加回数を 2 回削減することができた．

図 9 に選択されたサンプルと識別境界の可視化したものを示す．初期の入力サンプルと識別境界から従来法 (Vote Entropy) と提案手法によりラベルを 1 回追加した際の入力サンプルと識別境界，2 回追加した際の入力サンプルと識別境界の実際に選択されたサンプルと識別境界を比較すると，提案手法では 2 回目のラベル追加で教師付き学習と同等の識別率になる．従来法では密度推定にばらつきがある曖昧な領域に存在する類似したサンプルが選択されるため，2 個追加した場合でも識別境界が大きく変化しない．提案手法では密度推定のばらつきによる条件を用いることで，ラベル付きサンプルが周囲に存在しない領域のサンプルも追加することができる．そのため，類似したサンプルの選択を抑制し従来法と比べて識別境界が大きく変化する．

4.3 考察

図 10 に示す従来法 (Vote Entropy) と提案手法よりラベルを 1 回追加した際と，2 回追加した際のラベル伝播の結果を比較する．従来法と比較すると，提案手法のラベルの伝播の結果が教師付き学習に用いる入力サンプルに近いことが確認できる．ラベル伝播はラベル付きサンプルからの距離が遠いほどラベルの伝播精度が低下する．そのため，ラベル付きサンプル付近の識別精度は高く，ラベル付きサンプルが存在しない領域の識別精度は低くなる．類似したサンプルの選択を抑制する提案手法では，ラベルの存在し

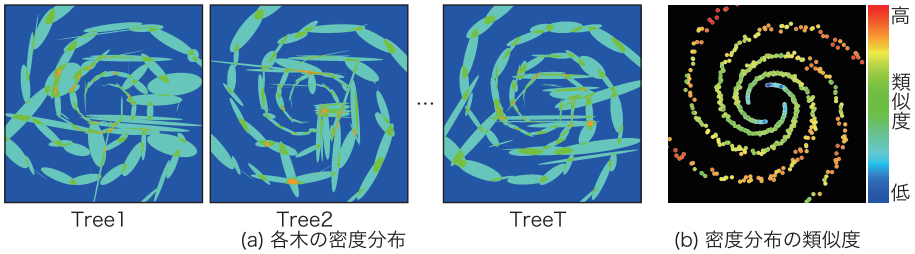


図 6 各木の密度分布とその類似度
Fig. 6 Density distribution of each tree, its similarity

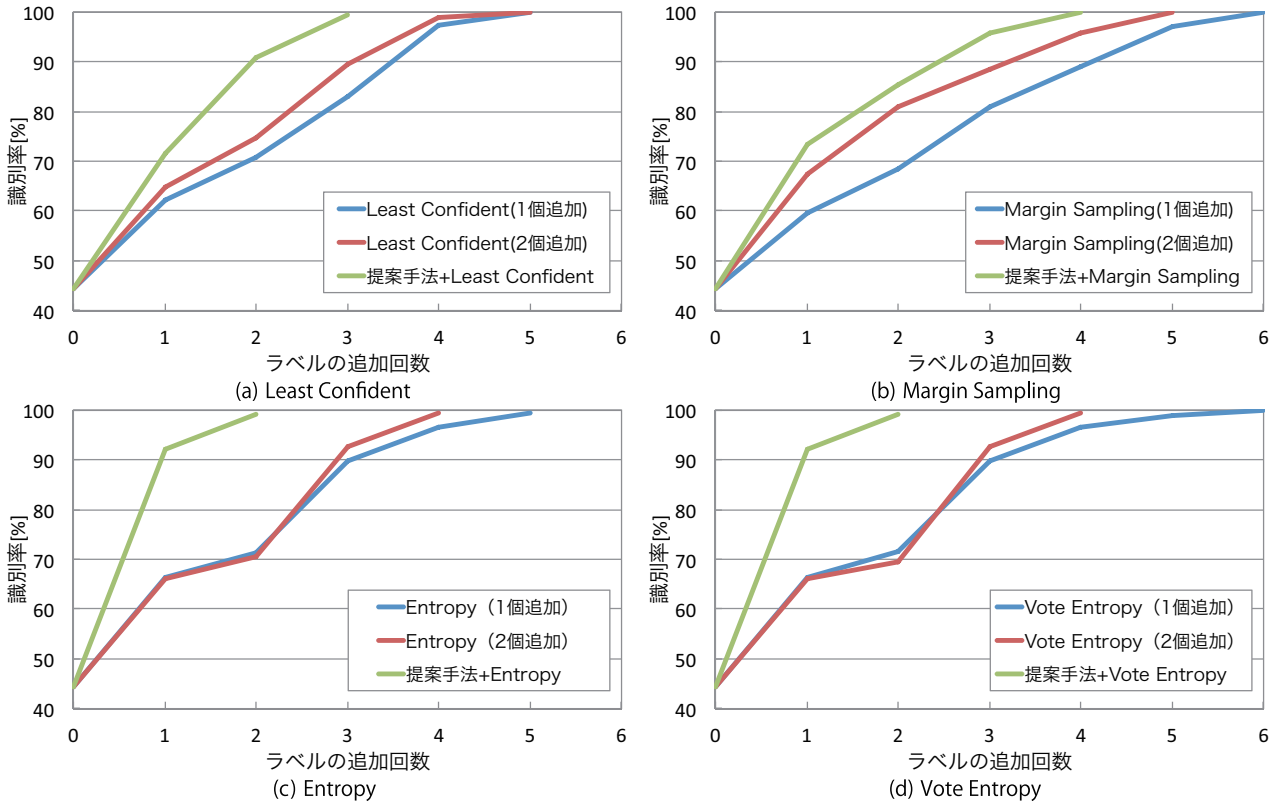


図 8 各手法によるラベルの追加回数の比較
Fig. 8 Comparison of the number of labeling times by each method

ないような領域のサンプルにラベルを追加することができるため，ラベル伝播の結果にも良い影響を与える．**表 1** に示す従来法と提案手法の各木のラベル伝播の正解率の平均からも，提案手法の有効性が確認できる．これらのことから，提案手法では少ない追加回数で目標である教師付き学習と同等の識別率を得ることができた．

5. おわりに

本研究では，密度分布の類似度を考慮したサンプル選択法を提案した．提案手法によるサンプル選択法を導入することで能動学習における追加回数を削減することができた．今後は大規模なデータセットや，高次元なデータセットに提案手法を適用し，提案手法の有効性について検証する予定である．

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*, CVPR (2009).
[2] A. Sorokin, D. Forsythi: *IUtility data annotation with Amazon Mechanical Turk*, CVPR (2008).
[3] Settles, Burr: *Active Learning Literature Survey*, Computer Sciences Technical Report1648, University of Wisconsin–Madison, (2007).
[4] Xiaojin Zhu: *Semi-Supervised Learning Tutorial*, Department of Computer Sciences, University of Wisconsin, (2009).
[5] L. Breiman: *Random Forests*, Machine Learning, vol. 45, pp. 5-32, (2001).
[6] A. Criminisi, J. Shotton, and E. Konukoglu: *Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning*, Foundations and Trends in Computer Graphics and Vision, vol. 7, no.2-3, pp. 81-227, (2012).

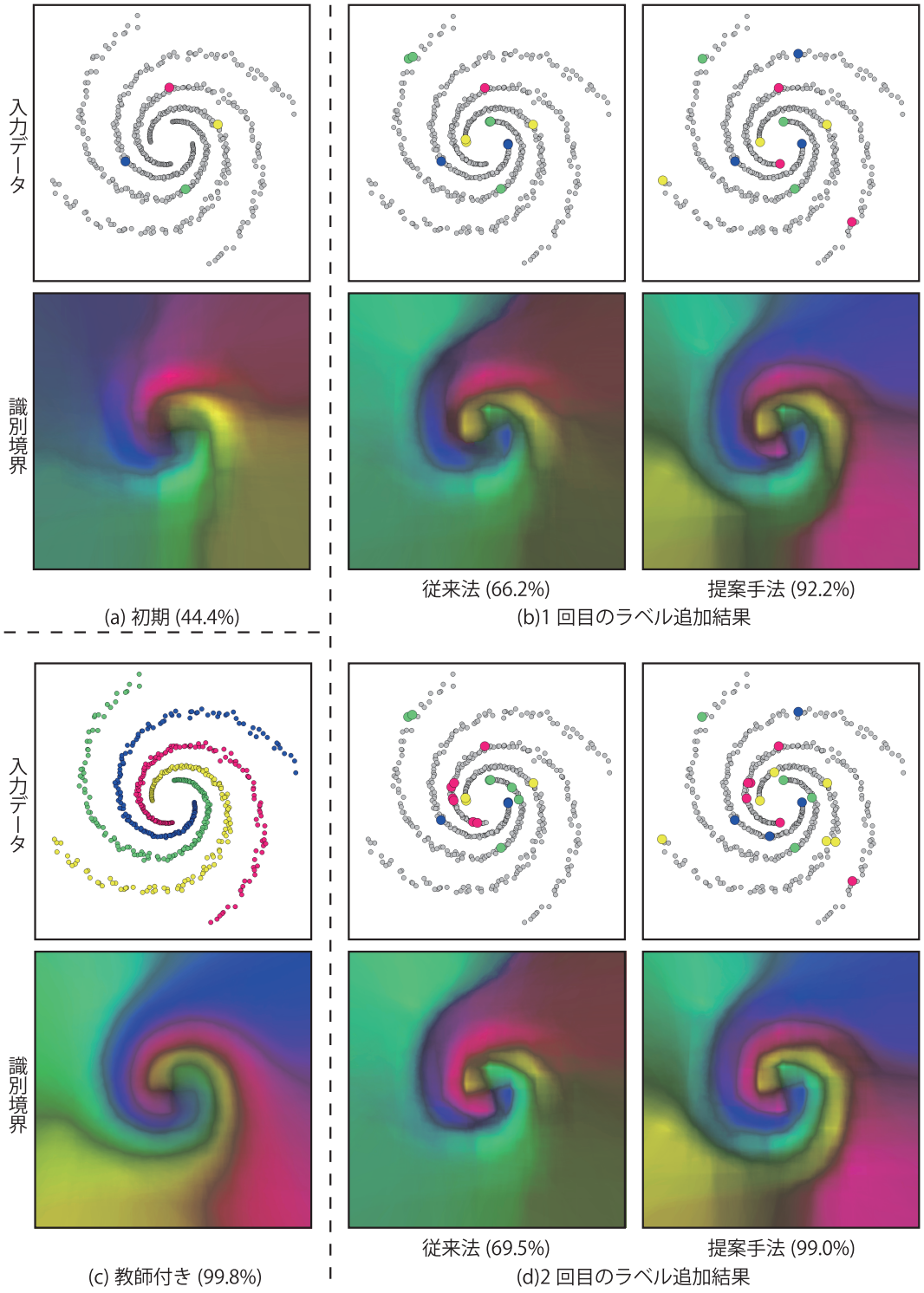


図 9 選択されたサンプルと識別境界の可視化

Fig. 9 Visualization of the selected samples and discrimination boundary

表 1 ラベルの伝播精度

Table 1 Propagation rate

	従来法		提案手法	
追加回数	伝播の正解率 (%)	識別率 (%)	伝播の正解率 (%)	識別率 (%)
1	54.1	66.2	72.0	92.2
2	66.1	69.5	85.9	98.0

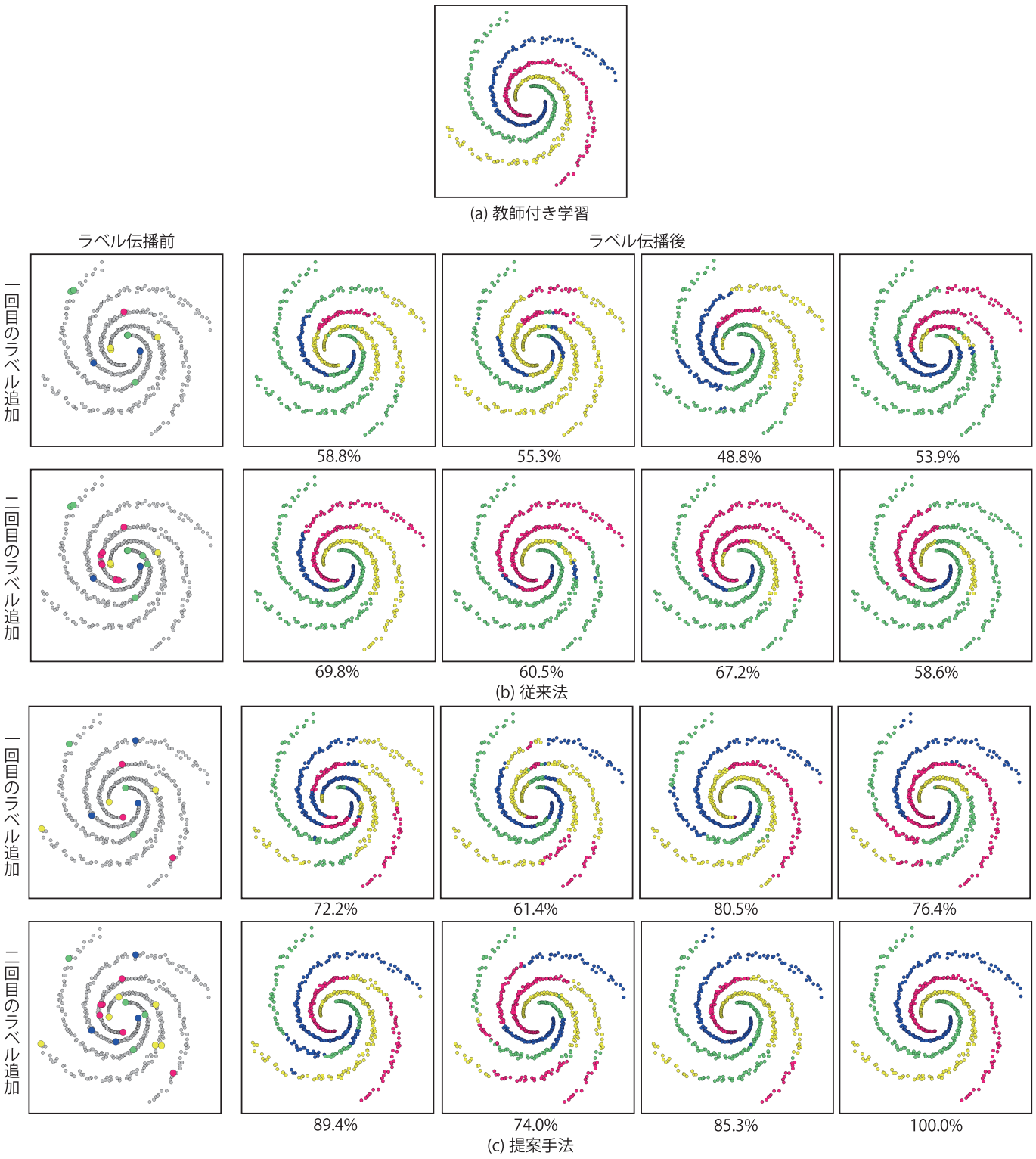


図 10 各木のラベル伝播結果の比較

Fig. 10 Comparison of propagation results by each method

[7] D. Lewis and J. Catlett.: *Heterogeneous uncertainty sampling for supervised learning*, ICML, pp. 148–156, (1994).
[8] T. Scheffer, C. Decomain, and S.Wrobel: *Active Hidden Markov Models for Information Extraction*, CAIDA, pp. 309-318, (2001).
[9] A. Holub, P. Perona, and M. Burl. : *Entropybased active learning for object recognition*, CVPR, Workshop on Online Learning for Classification, (2008).
[10] H. Seung, M. Opper, and H. Sompolinski: *Query by committee*, COLT, pp. 287-294, (1992).
[11] I. Dagan, and Sean P. Engelson.: "*Committeebased sampling for training probabilistic classifiers*, ICML. Vol. 95. (1995).