Random Forest を用いた能動学習における有効なサンプル選択

EP09137 村田 隆英　　指導教授：藤吉弘亘

1. はじめに

大量の学習データを用いて識別器を構築する際，どのサンプルに教師ラベルを付けることで，効率良く識別境界を決定できるかという問題がある．能動学習では，一度に追加するサンプルが少数の場合，類似したサンプルが選択され効率が悪いという問題がある．そこで，本研究では，Random Forest(RF) の密度推定を用いて適切なサンプルを選択してラベル付けすることで，効率良く識別境界を決定する手法を提案する．

2. 問題設定

能動学習 [1] では，図 1 に示すように，識別境界の定まらない領域にある曖昧なサンプルを選択してラベル付けを行い，再学習することでより良い識別境界を求める．従来，追加サンプルの選択には，複数の識別器の出力に統一性の無いサンプルを選択する Vote Entropy が利用されている．従来法では，サンプルの分布を考慮しないため，類似したサンプルを選択することがあり，効率が悪いという問題がある．

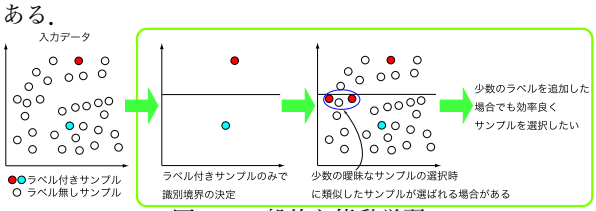


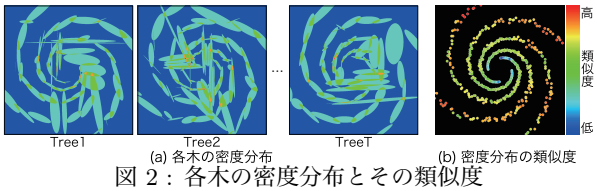
図 1：一般的な能動学習

3. 提案手法

本研究では，Density Forest による密度推定結果に着目したサンプル選択を行う．ラベルを追加すべきサンプルは，密度推定にばらつきがある曖昧な領域と，ばらつきは無いがラベル付きサンプルが周囲に存在しない領域にあると考え，この二つの領域に同時にラベルを追加するサンプル選択法を提案する．以下に提案手法の流れを述べる．

Step1：密度推定とラベル伝播による学習

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 $\mathcal{S}^{(u)}$ を用いて Density Forest を構築する．学習後の各密度木の密度分布を図 2(a) に示す．次に，各密度分布が連結している方向にラベルを伝播することで，ラベル無しサンプルにラベルを付与する．伝播結果により末端ノードのクラス分布を作成する．



(a) 各木の密度分布　　(b) 密度分布の類似度

図 2：各木の密度分布とその類似度

Step2：サンプルの曖昧さと密度分布の類似度の算出

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 は，シャノンの情報量を用いた JS-Divergence により式 (1) を用いて算出する．

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

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 2(b) に各サンプルの密度分布の類似度を示す．

Step3：密度分布の類似度を考慮したサンプルの選択

Vote Entropy の値 $VE(\boldsymbol{x}_i)$ と，密度分布の類似度 $D(\boldsymbol{x}_i)$ を用いてサンプルの選択を行う．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．

Step4：ラベルの追加

選択されたサンプルに対して人手によりラベル付けを行う．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．

Step5：ラベルの再伝播によるクラス分布の更新

ラベルの再伝播を行い，各木の末端ノードのクラス分布を更新する．提案手法では，決定木を再構築することはしない．**Step2**〜**Step5** を一定の条件に達するまで繰り返すことで識別境界を決定していく．

4. 評価実験

評価実験では，提案手法と従来法の Vote Entropy を比較する．

4.1. 実験概要

従来法と提案手法のラベルの追加回数を比較する．両手法において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．実験には，スパイラルデータ (2 次元) を使用する．RF のパラメータ木の数は 400 本，木の深さは 10 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．

4.2. 実験結果

図 3 に識別率が 99%に達するまでのラベルの追加回数とラベル再伝播後の識別率を示す．提案手法は，従来法よりラベルの追加回数を削減することができた．

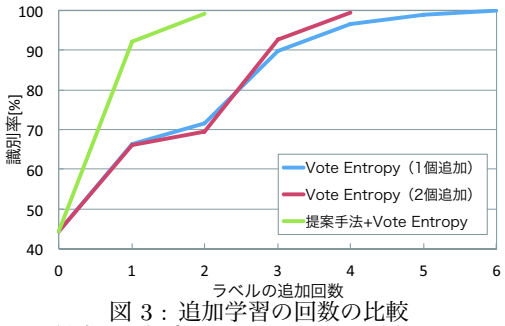


図 3：追加学習の回数の比較

図 4 に，従来法と提案手法よりラベルを追加 (2 回) した際の入力サンプルと識別境界を示す．従来法は，密度推定にばらつきがある曖昧な領域の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．提案手法は，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を追加することができるため，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．これにより，少ない追加回数で高い識別率を得ることができる．

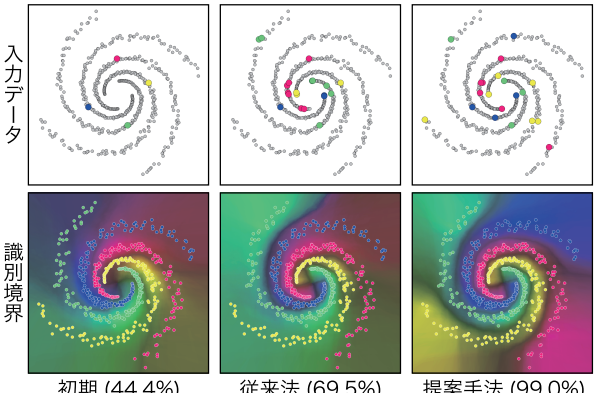


図 4：各手法の入力データと識別境界

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に提案手法を適用する予定である．

参考文献

[1] B. Settles,“Active Learning Literature Survey”, Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.