解釈性を有するアンサンブル識別器に関する一考察

1X17C139-1　良川太河
指導教員　後藤正幸

1. 研究背景・目的

機械学習分野において，ラベル分類の手法の一つである決定木は，人間が理解しやすい木構造で表現された解釈性の高い識別器である．しかし，有限の学習データに対して過学習を起こしやすく，十分なデータ数がない場合，予測精度が悪化する恐れがある．一方，Random Forest[1]（以下，RF）などのアンサンブル識別器は過学習を防ぎ，高い予測精度を示す手法であるが，複数の決定木を合成しているため，解釈性が失われてしまう．そのため，RF などのアンサンブル識別器に近い予測性能を維持しつつ単一の決定木を学習できれば，予測性能と解釈性の両面から大変有益なモデルとなる．

この問題に対して，アンサンブル識別器を活用して学習データを増やすことで，予測精度が高い単一の決定木を学習する手法が研究されている．その代表的な手法として，Born Again Trees[2](以下，BATrees) が挙げられる．BATrees は，学習データの特徴量にランダムノイズを加えたデータを大量に生成し，これらを学習済みのアンサンブル識別器に入力してラベル付けを行うことで，追加の学習データを大量に生成する手法である．しかし，BATrees のデータ生成法では膨大な計算量が必要となる上に，対象データの分布から外れたデータも多数生成してしまうため，学習した決定木が複雑になる傾向がある．そのため，どの識別規則が対象データの識別に寄与するかがわかりづらく，解釈性が低下してしまう．

そこで本研究では，生成モデルを用いて対象データの分布に従うデータを少ない計算量で生成することで，予測精度を維持しつつ，なるべくシンプルな決定木を学習する方法を提案する．具体的には，深層ニューラルネットワークに基づく生成モデルである Autoencoder[3](以下，AE) で得られる潜在表現データに対して，オーバーサンプリング手法の一つである Synthetic Minority Over-sampling Technique[4](以下，SMOTE) を適用する．SMOTE により，潜在表現データを大量に合成し，復元することで対象データの分布に従うデータを少ない計算量で生成する．これらと元の学習データを用いることにより，BATrees よりもアンサンブル識別器と近い予測精度を持つ単一の決定木の学習が可能となり，予測精度と解釈性を兼ね備えたモデルが期待できる．最後に，手書き文字のデータセットである MNIST を用いた評価実験を行い，提案手法の有効性を示す．

2. 準備

2.1. 問題設定

本研究では，特徴量 $\boldsymbol{x}$ からラベル y の予測を行う問題を扱う．n 個の学習データ集合を $(\boldsymbol{X}^n, \boldsymbol{Y}^n) = \{(\boldsymbol{x}_i, y_i)\}^n$ とし，i 番目のデータの特徴量を $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{iM})^\top$，ラベルを $y_i \in \{1, \cdots, L\}$ とする．n 個の学習データ $(\boldsymbol{X}^n, \boldsymbol{Y}^n)$ により学習した決定木を用いて未知のデータの特徴量 $\boldsymbol{x}$ からラベル y を予測する．このとき，図 1 に示すような決定木において，新たに与えられた $\boldsymbol{x}$ は対応するリンクを伝い，最終的に辿り着いたノードに該当するラベル y で予測を行う．

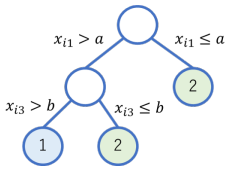

図 1: 決定木のモデル構造

2.2. Born Again Trees

BATrees では，学習データ $\boldsymbol{x}_i \in \boldsymbol{X}^n$ にノイズを加えた大量の特徴量 $\tilde{\boldsymbol{x}}$ に対して，学習済みのアンサンブル識別器を用いてラベル付けを行い，追加の学習データを生成する．データの生成過程は，大きく二つのステップに分けられる．

Step.A1 学習データ $\boldsymbol{x}_i \in \boldsymbol{X}^n$ の要素を一部変更した大量の $\tilde{\boldsymbol{x}}$ を生成する．

Step.A2 生成した $\tilde{\boldsymbol{x}}$ に対するラベル $\tilde{y}$ を付与する．

Step.A1 では，初めに学習データ $\boldsymbol{X}^n$ の中からランダムに $\boldsymbol{x}_i$ を一つ選択する．その後，$\boldsymbol{x}_i$ の各要素 x_{ij} に対して，一定の変更率 c で要素を $x_{lj}(l \in \{1, \cdots, n\})$ に変更し，この作業を終えた特徴量を $\tilde{\boldsymbol{x}}$ とする．Step.A2 では，得られた $\tilde{\boldsymbol{x}}$ に対して，事前に学習したアンサンブル識別器を用いて正解ラベル $\tilde{y}$ の付与を行う．これらの作業により，元々の n 個の学習データに対し，$N(\gg n)$ 個の追加の学習データ $(\tilde{\boldsymbol{X}}^N, \tilde{\boldsymbol{Y}}^N)$ を得る．これらと元の学習データ $(\boldsymbol{X}^n, \boldsymbol{Y}^n)$ を用いて単一の決定木を学習することで，十分な数の学習データを確保できるため，決定木の解釈性を保ちつつ，元の学習データ $(\boldsymbol{X}^n, \boldsymbol{Y}^n)$ のみを用いて学習したモデルより予測精度が高いモデルの学習が可能となる．

2.3. Autoencoder

AE は深層ニューラルネットワークに基づく生成モデルの一つであり，エンコーダとデコーダから構成されている．AE のモデル概要を図 2 に示す．AE は，エンコーダにより入力データから重要な特徴を抽出し，低次元の潜在表現 $\boldsymbol{Z}^n$ に変換した後，デコーダで入力データと同じ特徴量を出力するように復元処理を行うモデルである．また，AE におけるネットワークのパラメータは，学習データ $\boldsymbol{X}^n$ を用いて，入力データとそれに対応する出力との二乗誤差損失が最小となるように，誤差逆伝播法により学習される．

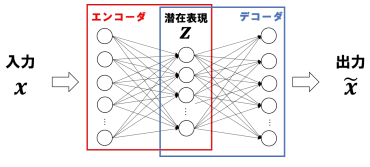

図 2: AE のモデル概要

2.4. SMOTE

SMOTE は，オーバーサンプリング手法の一つであり，以下の手順で対象ラベルのデータを生成する．

Step.B1 対象ラベルのデータ $\boldsymbol{w}_i$ をランダムに一つ選択し，k 個の近傍データ $\{\boldsymbol{w}_{i1}, \cdots, \boldsymbol{w}_{ik}\}$ を用意する．

Step.B2 $\{\boldsymbol{w}_{i1}, \cdots, \boldsymbol{w}_{ik}\}$ から，ランダムに $\boldsymbol{w}_{im}$ を一つ選択し，$\boldsymbol{w}_i$ と $\boldsymbol{w}_{im}$ の内挿または外挿で新しいデータを生成する．

3. 提案手法

3.1. 着想

本来，モデル学習のために新たに追加する学習データは，対象データの分布 $p(\boldsymbol{x})$ に従っていることが望ましい．しかし，BATrees では学習データの特徴量にノイズを乗せることで追加の学習データ $\tilde{\boldsymbol{x}}$ を生成しているため，$p(\boldsymbol{x})$ から外れたデータも多数生成してしまう．そのため，これらのデータを学習した決定木では，対象データの識別に寄与しないノードが増え，木が必要以上に複雑になってしまう．加えて，データ $\boldsymbol{x}_i$ の各要素 x_{ij} に対して処理を行っているため，膨大な計算量が必要となる．

一般に対象データの分布 $p(\boldsymbol{x})$ は複雑で，高次元な空間上に点在した疎なデータであるため，$\boldsymbol{x}$ の空間上の単純な合成手法で良質のデータを生成することは難しい．一方，深層ニューラルネットワークモデルの一つである AE では，対象データをより低次元で表現した潜在表現 $\boldsymbol{Z}^n$ を獲得し，$\boldsymbol{Z}^n$ にノイズを加え復元することで，新たなデータを生成する．ただし，単純なノイズを加えても，$p(\boldsymbol{x})$ に従うデータを生成できるかどうかは不明である．そこで提案手法では，AE のエンコーダで得られた潜在表現 $\boldsymbol{Z}^n$ に SMOTE を活用することで，潜在表現の分布 $p(\boldsymbol{Z})$ に従う，新たな潜在表現 $\tilde{\boldsymbol{Z}}^N$ を合成する．得られた潜在表現 $\tilde{\boldsymbol{Z}}^N$ を AE のデコーダで復元することで，各データ $\boldsymbol{x}_i$ の中間的な特徴を持ち，$p(\boldsymbol{x})$ に従う追加の学習データ $\tilde{\boldsymbol{X}}^N$ を得ることができる．さらに，RF を用いて $\tilde{\boldsymbol{X}}^N$ にラベル $\tilde{\boldsymbol{Y}}^N$ を付与する．これらの追加データ $(\tilde{\boldsymbol{X}}^N, \tilde{\boldsymbol{Y}}^N)$ と元の学習データ $(\boldsymbol{X}^n, \boldsymbol{Y}^n)$ を用いることで，BATrees と同程度の精度を持ち，より解釈性の高い単一の決定木を学習する．加えて，提案手法ではデータ $\boldsymbol{x}_i$ 単位でデータを生成するため，BATrees に比べ非常に少ない計算量でデータを生成することが可能である．

3.2. データの生成方法

提案手法では，下記の手順で $(\tilde{\boldsymbol{X}}^N, \tilde{\boldsymbol{Y}}^N)$ を生成する．

Step.C1 学習データ $\boldsymbol{X}^n$ により AE を学習する．

Step.C2 AE のエンコーダに学習データ $\boldsymbol{X}^n$ を入力し，学習データの潜在表現 $\boldsymbol{Z}^n$ を得る．

Step.C3 潜在表現 $\boldsymbol{Z}^n$ に対し，ラベルごとに SMOTE を適用することで新たな潜在表現 $\tilde{\boldsymbol{Z}}^N$ を生成する．

Step.C4 得られた $\tilde{\boldsymbol{Z}}^N$ を AE のデコーダに入力し，その出力を追加の学習データ $\tilde{\boldsymbol{X}}^N$ とする．

Step.C5 RF を用いて $\tilde{\boldsymbol{X}}^N$ のラベル $\tilde{\boldsymbol{Y}}^N$ を得る．

上記の過程で得た追加の学習データ $(\tilde{\boldsymbol{X}}^N, \tilde{\boldsymbol{Y}}^N)$ と元の学習データ $(\boldsymbol{X}^n, \boldsymbol{Y}^n)$ を用いて，単一の決定木を学習する．

4. 評価実験

4.1. 実験条件

実験データには，28 × 28 ピクセルの計 784 次元の特徴量を持ち，0 から 9 までの 10 種類を含む手書き数字の画像データセットである MNIST を用いる．MNIST の各要素は 0 から 255 までのグレースケールの色表現の値を持つが，本実験では簡略化のため，各要素を 0-1 に正規化し，閾値を 0.5 として 0 と 1 に離散化した．本実験では，学習データを 8,000 個，テストデータを 2,000 個として，5 分割交差検証を行った．また，評価指標としてテストデータに対する信頼区間 95%の予測精度，得られたモデルの解釈性を表すノード数及びデータ生成にかかる所要時間を用いる．比較手法には，通常の決定木の学習法である CART と，従来手法である BATrees を用いる．学習データに追加するデータとして，BATrees では 16 万個 (160K)，提案手法では 16 万個 (160K) と 100 万個 (1M) の二通りの生成を行い，CART で決定木を学習した．また，アンサンブル識別器として RF を適用してラベル付けを行った．なお，各ノードにおける最小データ数は 10，BATrees の変更率は c=0.25 とし，提案手法で使用する AE には 3 層のニューラルネットワークを用いる．また，入力データの次元数を 784，潜在表現の次元数を 32 とした．

4.2. 結果と考察

表 1 に，各手法の識別精度，学習した決定木のノード数及びデータ生成にかかった所要時間を示す．

表 1: MNIST における学習したモデルの評価

	識別精度 (信頼区間)	ノード数 (個)	所要時間 (s)
CART	0.7600 (±0.0182)	—	—
BATrees(160K)	0.8140 (±0.0086)	17,341	8,159.1
提案手法 (160K)	0.8142 (±0.0114)	**5,233**	**136.8**
提案手法 (1M)	**0.8271** (±0.0120)	16,783	185.7

表 1 より，提案手法は BATrees に比べデータ生成のコストが低く，少ない所要時間で 100 万枚の追加データを生成することができる．これらを用いて学習した提案手法 (1M) は，BATrees に比べて識別精度で有意な改善が見られる．また，提案手法はいずれの場合も BATrees より少ないノード数で木を学習できているため，解釈性が高い木を学習できていることもわかる．これは，提案手法が対象データの入力分布に従うデータを多く生成できていることで，識別に寄与しないノードが生成されにくくなったためと考えられる．

5. まとめと今後の課題

本研究では，AE を活用して BATrees よりも解釈性の高いモデルをより少ない計算量で作成する方法を提案し，MNIST を用いた実験により手法の有効性を示した．今後の課題としては，様々なデータセットに対する応用などが挙げられる．

参考文献

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

[2] Breiman.L, “Born Again Trees,” *Technical Report, University of California Berkeley*, 1996.

[3] Hinton.G, “ Reducing the Dimensionality of Data with Neural Networks,” *Science (American Association for the Advancement of Science)*, vol. 313, no. 5786, pp. 504–507, 2006.

[4] Chawla, Bowyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *The Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.