7Q-08

# Correct and Smooth を用いたアンサンブル手法に関する一考察

石川　悠樹† 三川　健太‡　二宮　洋†

湘南工科大学工学部情報工学科†　　東京都市大学メディア情報学部情報システム学科‡

## 1. 研究背景

近年，グラフ構造を持つデータセットに対する機械学習手法が注目されており，その一手法として Correct and Smooth（以下，C&S）と呼ばれる手法が存在する [1]. C&S は軟判定が可能な任意の識別器を用いた初期予測を行い，それを基に初期予測の誤差の伝搬と，ラベルの伝搬に基づくテストデータの分類を行うことで，少ない学習コストで高性能な分類が可能とする．前述の通り，C&S はグラフ構造を持つデータに対する手法であるが，これをグラフ構造を持たないデータに適用するための手法が存在する [2].

本研究では，従来手法 [2]における分類精度向上を目的に，そのアンサンブル手法の検討を行う．具体的には，C&S が初期予測，誤差の伝搬，ラベルの伝搬を用いた最終予測の 3 ステップから構成されていることに着目し，各ステップにおいてアンサンブルを行うための方法論の提案を行う．新聞記事データを用いた文書分類実験により，提案手法の有効性を示す.

## 2. 準備

### 2. 1 問題設定

$n = |V|$のノードを持つ無向グラフを$G = (V, E)$とし，各ノードの特徴量を成分として持つ行列を$X \in R^{n \times p}$とする．グラフ$G$における隣接行列を$A$，次数行列を$D$とし，正規化隣接行列$S$を$S = D^{-1/2}AD^{-1/2}$として定義する．分類のため，ノード集合$V$をラベル付きノード$L$とラベルなしノード$U$に分割する．ラベル行列$Y$はクラス数を$c$としたもとで one-hot-encoding で表されているものとし，$Y \in R^{n \times c}$と定義される．また，ラベル付きノードを訓練ノード$L_t$と検証ノード$L_v$に分割する．これらの情報を用いて，ラベルの付与されていないノード$j \in U$に対する分類を行う．

Ensemble method using Correct and Smooth
†Yuki Ishikawa, ‡Kenta Mikawa, † Hiroshi Ninomiya
†Department of Information Science,
Shonan Institute of Technology
‡Department of Information Systems,
Tokyo City University

### 2. 2 Correct and Smooth

C&S はグラフ構造に対応したトランスダクティブノード分類の手法であり，1)初期予測，2)学習データを用いた誤差伝搬，3)ラベル伝搬を用いた最終予測の３ステップにより構成される．

1)ではグラフ構造を利用せずに学習データの予測を行う．C&S の特徴として，軟判定を行う分類モデルであれば任意の分類器を使用することが可能である．いま，初期予測によって得られた予測ラベルを$Z \in R^{n \times c}$とし，$Z$の各行はデータの各クラスへの所属確率を示すものとする．初期予測における誤差が発生した場合，隣接するノードにおいても同様の誤差が生じる可能性が高いという仮定のもと，2)ではラベルと予測値との残差の伝播を行う．このため，誤差行列$E \in R^{n \times c}$を以下のように定義する．

$$E_{L_t} = Z_{L_t} - Y_{L_t},\ E_{L_v} = 0,\ E_U = 0 \tag{1}$$

すなわち，誤差行列$E$はラベルが存在する成分には予測値と真の値との差分を，それ以外は 0 をその成分とする行列となる．これに対し，ラベル拡散法を用いることで誤差行列を

$$E^{(t+1)} = (1-\alpha)E + \alpha SE^{(t)} \tag{2}$$

として更新する．ただし，$t$は繰り返し回数とし，$\alpha = 1/(1+\mu), E^{(0)} = E$ とする．また，式(2)が収束した結果を$\hat{E}$とする．これにより，初期予測に誤差伝搬を行った結果を補正した予測値$Z^{(r)} = Z + \hat{E}$を算出する．3)では，グラフ内の隣接するノード同士は類似したラベルを持つことが多いという考えのもと，ラベル伝播を用いて予測値をさらに平滑化する．まず，正確なラベル伝播を行うため，予測値行列$G \in R^{n \times c}$を

$$G_{L_t} = Y_{Lt},\ G_{L_{v,\ U}} = Z^{(r)}_{L_{v,\ U}} \tag{3}$$

とする．そして，収束するまで

$$G^{(t+1)} = (1-\alpha)G + \alpha SG^{(t)} \tag{4}$$

を繰り返す．これによって得られた最終予測値$\hat{Y}$によって各ノードを分類する．

### 2. 3 グラフ構造を持たないデータセットへの C&S の適用

C&S はグラフ構造を持つデータセットに対しする利用を前提としており，その実行には隣接

行列が必要となる．これに対し，従来手法 [2]では教師あり学習の問題設定のもと，学習データに付与されているラベル情報を活用することで擬似的な隣接行列を構成し，分類問題に適用している．従来手法における分類は以下の手順により行われる．

Step 1) データ間の距離行列をもとに $k$ 近傍点集合を作成し， 近傍点間でエッジを構成する．
Step 2) 隣接データ間で類似度を求め，その値を重みとした隣接行列を作成する．
Step 3) エッジを類似度により重みづけた重み付き次数行列を作成し，それを用いた正規化隣接行列を作成する．
Step 4) 求めた隣接行列をもとに，C&S を用いた誤差伝搬，ならびにデータの分類を行う．

## 3. 提案手法

本研究では，従来手法の分類精度向上を目的にそのアンサンブルを行う．アンサンブルを行う際には，初期予測の時点で複数の識別器をアンサンブルする方法（以下，手法 1），最終予測の際にアンサンブルを行う，すなわち複数の C&S モデルをアンサンブルすることで最終予測を行う方法（以下，手法 3）が考えられる．本研究では，これらに加えて C&S における誤差伝搬を複数のモデルにより行い，これにより得られた結果をアンサンブルするための方法（以下，手法 2）について提案を行う．

手法 2 では，C&S が式(2), (4)を利用して誤差およびラベル情報の伝搬を行うことに着目する．いま，複数の識別器を用いた初期予測値を算出し，各初期予測値を利用して誤差伝搬を行うことで，予測値$Z^{(r)}$を複数算出する．これにより得られた結果をアンサンブルする．さらに，この予測値をもとに式(4)にあるようなラベル伝搬を行うことで最終予測値$\hat{Y}$の算出を行う．

手法 2 の手順について以下に示す．

Step 1) 従来手法の Step 1) - Step 3)と同様に，正規化隣接行列を算出する．
Step 2) 複数の識別器を用いた初期予測値を計算．
Step 3) 得られた各初期予測値に対して式(2)を用いて誤差の伝搬を行う．
Step 4) Step 3)で得られた結果をアンサンブルする．
Step 5) Step 4)でアンサンブルした結果に対し，式(4)を用いることで平滑化を行い最終予測値$\hat{Y}$を算出，テストデータの分類を行う．

## 4. 実験

提案手法の有効性を示すため，2017 年発行の毎日新聞記事を対象とした分類実験を行った．

### 4. 1 実験条件

データセットは 8 カテゴリ，合計 54545 件のデータからなり，学習データ 13636 件，検証用データ 13636 件，テストデータ 27273 件とした．特徴量として TF-IDF を用い，NMF により 100 次元まで削減するものとした．アンサンブルする識別器としてロジスティック回帰，サポートベクターマシン，ランダムフォレストを用い，最終予測は多数決を用いるものとした．比較手法は前述の 3 種類の識別器を用いたスタッキングアンサンブルとし，メタモデルとしてロジスティック回帰を用いた．その他，C&S 等のハイパーパラメータは各手法で最適となるように設定した．

### 4. 2 実験結果

表 1 にそれぞれの識別器を用いた際の初期予測時の精度，ならびに従来手法を用いた際の分類精度を，表 2 に提案手法 1 ~ 3 と比較手法の精度を示す．

表 1: 従来手法の精度

| 識別器 | ロジスティック回帰 | SVM | ランダムフォレスト |
|---|---|---|---|
| 初期予測 | 0.6780 | 0.5231 | 0.6705 |
| C&S 後 | 0.7204 | 0.5825 | 0.7111 |

表 2: 提案手法と比較手法の精度

| | 手法 1 | 手法 2 | 手法 3 | 比較手法 |
|---|---|---|---|---|
| 精度 | 0.7542 | **0.7722** | 0.7325 | 0.7075 |

表 1 と表 2 を比較すると，従来手法で最も精度が高いロジスティック回帰に対し，すべての提案手法で精度が優れていることが確認できる．また，比較手法に対しても，提案手法を用いることでより高性能な分類が可能となることが明らかとなった．これらより，提案手法の有効性を示すことができた．

## 5. まとめと今後の課題

本研究では分類問題を対象とした C&S に対し，アンサンブルを用いた精度向上法を提案した．実験結果より，提案手法は従来手法と比較して高い分類精度を達成可能であることを示した．

今後の課題として，C&S の特徴をさらに活かしたアンサンブル法の構築が挙げられる．

## 参考文献

[1] H. Qian, et al. "Combining label propagation and simple models out-performs graph neural networks," *Proc. 9th International Conference on Learning Representations (ICLR 2021)*, 2021.
[2] 石川 悠樹，三川 健太，”Correct and Smooth の分類問題への適用に関する一考察,” 日本経営工学会 2022 年春期大会予稿集, 2022.