遺伝的アルゴリズムによる機械学習における疑似ラベル生成手法の提案

第 1 グループ　細川 岳大

1. はじめに

近年，ラベル付きデータ生成に伴うコストの観点から分類問題において半教師あり学習 (Semi-Supervised Learning: SSL) と呼ばれる学習データ全体のうち一部にのみラベル付けされた状態で学習を行う手法が盛んに研究されている．しかし，当然のことながら各ラベルに対するデータ数が少ないほど精度が安定しなくなるという問題も報告されている．一方で，ラベルなしデータへのラベル付けタスクは一種の組合せ最適化問題と考えることができる．

そこで本研究では，組合せ最適化遺伝的アルゴリズムを用いてデータにラベルを付与することでラベル付きデータが少ない場合における半教師あり学習の頑健性を高める手法を提案する．

2. 要素技術

2.1. Contrastive Learning

Contrastive Learning (CL) とは，特徴表現を獲得するための自己教師あり学習のひとつである．一つの画像から得られる特徴表現が画像変換によって画像の持つ意味が大きく変わらない変形を獲得することができる．

2.2. SimCLR

Simple framework for Contrastive Learning of visual Representation(SimCLR)[1] は SSL の一つであり，CL により特徴抽出器を学習したうえで，少量のラベル付きデータによる Fine tuning によって分類器を獲得する手法である．

2.3. 遺伝的アルゴリズム

遺伝的アルゴリズム (Genetic Algorithm: GA) は生物の進化を模倣した最適解探索アルゴリズムである．選択，交叉，突然変異の 3 つの操作によって最適解を探索することができる．

3. 提案手法

本研究では， CIFAR-10 においてラベルありとラベルなしに分割し，ラベルなしデータの一部に対するラベルを GA によって探索する．GA の個体について各遺伝子はラベルデータのいずれかの整数値とし，取り出されたラベルなしデータに対する疑似ラベルとする．以下に提案手法の概要を示す．

1. CL によって特徴抽出器を学習する．
2. GA の個体と対応するデータを学習データとし， Fine tuning を行い，ラベル付きデータに対する識別率を適応度として GA の探索を行う．
3. 探索された個体とラベル付きデータを学習データとし，Fine tuning を行い，テストデータを識別する．

4. 数値実験

4.1. 実験方法

データセットとして 10 クラス識別問題である CIFAR-10 を用いた．学習に 50000 件，テストに 10000 件をそれぞれ使用し，ラベル付きデータ及びランダムに選ばれた疑似ラベルを付与する対象であるラベルなしデータはどちらも学習データから選んだ．CL の学習について，特徴抽出器には Resnet-18 を用い，学習データ 50000 件を，500 epoch 学習した．また分類器は MLP を使用した．GA の適応度計算については，30 epoch 学習した．

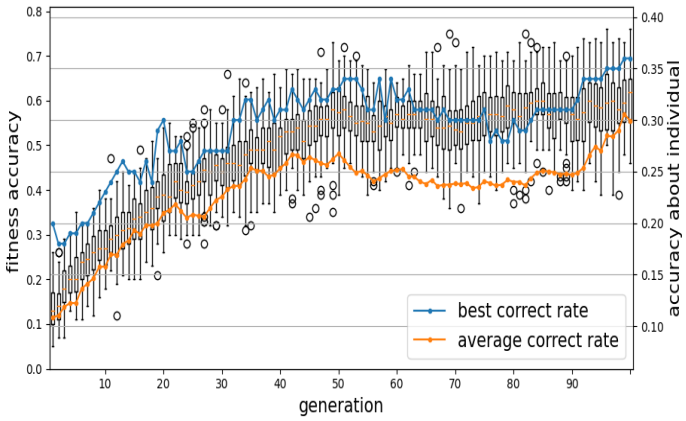

図 1: 各世代に対する適応度と疑似ラベルの正答率

4.2. 結果と考察

表 1 と図 1 に結果の一例を示す．また，図 1 の横軸に世代を，縦軸は箱ひげ図が適応度，折れ線グラフが各個体の疑似ラベルに対する正解ラベルの割合である． 提案手法が最も低い精度となった一方で，GA の探索については進んでいるといえ，疑似ラベルに対する真の正答率は 37%まで上がっていることが確認できた．

考察として GA の探索において今回の設定では多様性を保つことができず，局所解に陥ってしまったのではないかと考えられる．また，ラベルの多くが誤ラベルのため train_accuracy が学習において 80%以上にならず各個体に対しての正確な適応度が求められなかった可能性がある．

表 1: テスト識別率結果

生成ラベル	テスト識別率
baseline モデルによる疑似ラベル	0.784
提案手法による疑似ラベル	0.674
正解ラベル	0.822
baseline (ラベル付きデータのみ)	0.772

5. まとめと今後の課題

本研究では，GA によるラベルなしデータに対する疑似ラベルの生成手法を提案した．

今後の課題として，GA の拡張手法や適応度の計算方法を工夫することでより良質な疑似ラベルを生成し識別率向上を目指すことが挙げられる．

参考文献

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.