GA による疑似ラベル探索を用いた半教師あり学習の検討

1　はじめに

近年，機械学習は様々な分野への応用がなされており，様々なデータセット，タスクにおいて良い性能を示している．しかし，新規のデータセットを制作するにあたり分類タスクではデータに対するラベル付けのコストが問題となっており，それを解決する手法として半教師あり学習 (Semi Supervised Learning:SSL) という少量のラベル付きデータからラベルなしデータに疑似ラベルを付与する手法が提案されており，研究が盛んに研究されている．特に今年発表された FixMatch[1] という手法ではラベル付きデータが各ラベル 1 枚である場合でもかなりの精度を示すことが報告されている．一方でラベル付きデータが少ないと精度のばらつきも非常に大きくなってしまう．

そこで本研究では，ラベル付きデータを遺伝的アルゴリズムで増やすことでラベル付きデータが非常に少ない場合における半教師あり学習の頑健性を高めることを目的とする．

2　要素技術

2.1　FixMatch

FixMatch[1] は SSL の一つである．図 1 に FixMatch 概略図を示す．

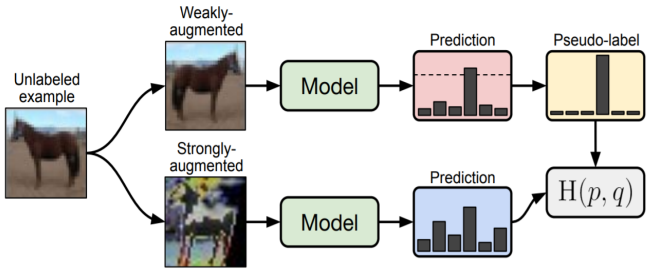

図 1: FixMatch[1] の概略図

この手法は Pseudo Label (疑似ラベル) と Consistency Regularization の統合した手法である．疑似ラベルはラベル付きデータから生成されたモデルに対しラベルなしデータを入力としたときの出力のうち最も確信度の高いラベルをそのデータのラベルとするものであり，Consistency Regularization は画像データに変換をかける前後において出力値が変化しないような制約をかける正則化手法である．

FixMtach におけるラベル付きデータのバッチサイズを B ，ラベルなしデータのバッチサイズを μB とする．$p_{\mathrm{m}}(y|x)$ を入力 x に対するモデルの出力，$\mathrm{H}(p,q)$ を確率分布 p,q に対する Cross Entropy Loss する．また，画像の弱変換，強変換をそれぞれ $\alpha(\cdot)$ ，初期 $\mathcal{A}(\cdot)$ とすると， FixMatch の loss は (5) 式となる．

$$l_{\mathrm{s}} = \frac{1}{B}\sum_{b=1}^{B}\mathrm{H}(p_b, p_{\mathrm{m}}(y|\alpha(x_b))) \quad (1)$$

$$q_b = p_{\mathrm{m}}(y|\alpha(u_b)) \quad (2)$$

$$\hat{q}_b = \mathrm{argmax}(q_b) \quad (3)$$

$$l_{\mathrm{u}} = \frac{1}{\mu B}\sum_{b=1}^{\mu B}\mathbb{1}(\max(q_b) \geq \tau)\mathrm{H}(\hat{q}_b, p_{\mathrm{m}}(y|\mathcal{A}(u_b))) \quad (4)$$

$$loss_{\mathrm{total}} = l_{\mathrm{s}} + \lambda l_{\mathrm{u}} \quad (5)$$

このとき τ は確信度の閾値であり λ はラベルなしデータにおける loss の重みである．1 式はラベル付きデータに対する loss であり，2 式は弱変換後のラベうなしデータに対する model の予測，3 式の $hat q_b$ は疑似ラベル，4 式はラベルなしデータに対する loss であり，1 式と比較して弱変換が強変換に，確信度が低いものにマスクをかける関数が組み込まれている違いがある．

2.2　遺伝的アルゴリズム

遺伝的アルゴリズム (Genetic Algorithm:GA)[2] とは生物の進化を模倣して適切なデータを見つけるアルゴリズムである．最小単位を遺伝子とし，探索するデータを遺伝子の集合である個体として表現する．各個体の適応度を計算し，個体の集まりである集団に対し選択，交叉，突然変異の三種類の操作を適用させ次の集団を作る，という操作を繰り返して適応度の高い個体を探索する．交叉の特性上，他のアルゴリズムより局所探索になりにくいが，一方で設定によっては初期収束を起こしてしまう．

3　データセット

データセットについて CIFAR-10 を用いた．CIFAR-10 は 6 万枚の画像からなるデータセットであり，各画像 32 × 32 pixel のカラー画像でそれぞれに {airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck} の 10 種類のラベルがついている．

4　提案手法

今回は FixMatch に GA を導入した手法について提案する．

1. データをラベル付きデータ (:D_{L})，ラベルなしデータ (:D_{NL})，探索データ (:D_{S})，テストデータ (:D_{T}) に分割する．
2. D_{L} の一部の D_{NL} を用いて FixMatch により model を学習させる
3. 生成された model から探索データに仮のラベルを付け，GA の初期個体のベースとし，手順 2 での未使用のラベルなしデータの error 率で各遺伝子座ごとに突然変異させて初期個体を得る．
4. 得られた初期個体から GA を回す．
5. 最終的に得られた個体を探索データに対するラベルとみなし，ラベル付きデータに追加し FixMatch で再学習しテストデータに対する精度を求める．

4.1　GA の設定

4.1.1　個体表現

各遺伝子はラベルなしデータに対する CIFAR-10 のラベルデータである 0 ～ 9 までのいずれかの整数値を持つ．全個体を通して各遺伝子座に対する探索データが共通であり一意に決まっているため，個体の遺伝子長は探索するデータ数となる．

4.1.2　選択

選択手法はエリート選択とトーナメント選択を用いた．エリート選択は前世代における最大の適応度を持つ個体を必ず次の世代に残す手法で本実験では毎世代 2 つ選択する．トーナメント選択はランダムに複数の個体を選びそのうちで最も高い適応度を持つものを選択する方法である．なお，ランダムに選ぶ個数をトーナメントサイズと呼び，本実験では 3 に設定した．

4.1.3　交叉

交叉手法は二点交叉を用いた．二点交叉とは，交叉する 2 つの個体を三分割し，それぞれの部分の遺伝子らを入れ替える手法である．

4.1.4　突然変異

突然変異はある遺伝子に対し他の対立遺伝子へとランダムに変更するものとした．また突然変異率について各遺伝子座に対し 5%でほかのラベルに変化するものとした．

5　数値実験

半教師あり学習のタスクとして，本実験では cifar10 の train_data 50000 枚のうちラベル付きデータを各ラベル 25 枚，計 250 枚のみを用いるタスクで実験した．また今回探索するデータについて，ランダムな選択ではあるものの，各正答ラベルが均等になるように選ばれている．表 1，2 に実験設定を示す．

6　実験結果および考察

図 2,3 に実験結果を示す．図 2 について箱ひげ図は各個体の適応度を示しており，折れ線グラフは各世代における探索データに対する正解ラベルの最大と平均の割合を示しており，共に左の数値に従っている．また横軸は世代数となっている．図 3 は縦軸が適応度，横軸が個体の正解ラベルに対する割合となっており，探索された全個体についての散布図である．

図 2 より最大正答数の増加は見られないが平均の正答数は上がっていることは確認できる．また，適応度としても収束しつつあり，これ以上の精度改善は困難であると思われる．また図からは読み取ることはできないが，各世代の適応度最高値は正答数が多いものだけに関わらず，特に正答数の低いものも選ばれることも多々あった．これはデータ数が非常に少ないことが大きな原因であることが考えら

表 1: FixMatch の設定

model	WideResNet16-2	
data set	CIFAR-10	
batch size	labeled	32
	unlabeled	32 × 7
optimizer	SGD(lr=0.1,momntum=0.9)	
事前学習		
train	labeled	100
	unlabeled	49650
val data	150	
num_iterations	2^15	
個体の適応度の評価		
train	search のみ	100
	unlabeled	49650
val data	250	
num_iterations	5000	
得られた個体の評価		
train	labeled+search	250+?
	unlabeled	49650
val data	10000	
num_iterations	2^16	

表 2: GA の設定

個体数	20
世代数	20
交叉率	1.0
突然変異率	0.05
labeled	250 枚
search	100 枚

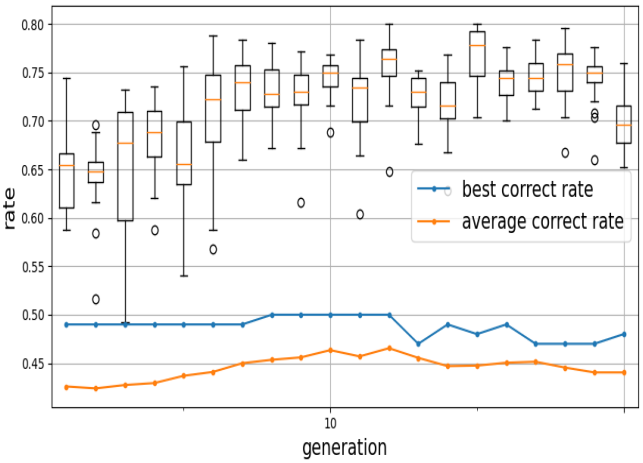

図 2: 実験の結果

れる．さらに 12 世代以降で最大及び平均の正答数が減っていることもわかる．これは過学習と同様に 250 枚という少ないデータに適合しすぎて汎化性が失われていると考えられる

図 3 から相関性はありそうではあるが，実際には学習が全く進んでいない．これは先にも述べた通りデータの少なさゆえに正答数に対して適応度に幅が出てしまっているからであることが分かる．

また表 3 に適応度の平均が最高であった 14 世代の個体を用いて再学習した結果を示す．採択される遺伝子はある遺伝子座を全個体通してみた時に最も多く出現した遺伝子とする．閾値はある遺伝子座において採択された遺伝子の出現数が全個体に対して占めていた割合に対するものである．

結果としてラベル付きデータのみを用いたものを超えることはできない結果となった．従来のシンプルな疑似ラベルを追加する手法 [3] において疑似ラベルの役割としてエントロピーの正則化が挙げられている．つまり FixMatch における正則化が強く，探索されたデータの疑似ラベルの正則化がうまく働かず，探索を阻害したと考えられる．

結論として，データが少ない状況において cifar10 といった画像データかつ 10 種類の分類タスクにおけるラベル付けは今回の設定の GA の学習でほとんど得られないことが分かった．

表 3: GA の設定

採択数	正答数	閾値	精度
0			0.868
100	46	なし	0.836
39	22	0.19	0.862
19	14	0.2	0.825

7　まとめと今後の課題

本研究から GA によるラベルなしデータのラベル付けを提案した．しかし，結果として GA で得られ

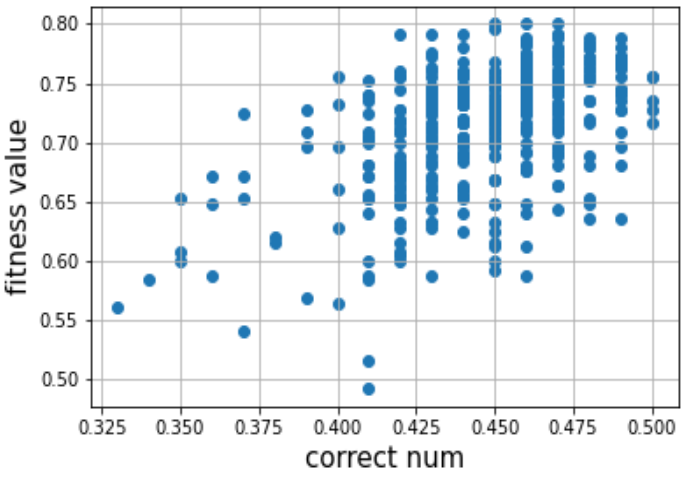

図 3: 実験の相関図

た疑似ラベルによって FixMatch の精度改善にはつなげることはできなかった．ただしデータとしての正答数の平均自体の改善はみられるため FixMatch の loss として GA のような動作を組み込むことができればと考えている．また，近年の半教師あり学習において自己教師あり学習を組み込んだもの [4, 5] がより成果を出しており，今後の課題としてそれらに GA が適用できないかということがあげられる．

参考文献

[1] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[2] Darrell Whitley. A genetic algorithm tutorial. *Statistics and computing*, 4(2):65–85, 1994.

[3] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.

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

[5] Xiao Wang, Daisuke Kihara, Jiebo Luo, and Guo-Jun Qi. Enaet: Self-trained ensemble autoencoding transformations for semi-supervised learning. *arXiv preprint arXiv:1911.09265*, 2019.