画像分類タスクのための能動学習における効果的なデータ選択方法に関する研究

1X19C111-8　益田恵里花
指導教員　後藤正幸

1. 研究背景と目的

近年，多くの企業が情報技術を活用したビジネスを展開する中で，大量のデータを蓄積している．そして，蓄積したデータを経営やマーケティングなど，様々な場面における意思決定に活用するために，機械学習を用いることが一般的になりつつある．

機械学習手法の中でも, 特に教師あり学習は有用なアプローチとして広く用いられている. 教師あり学習ではモデルを学習する際, 全ての学習データに正解ラベルが付与されている必要がある．中でもディープニューラルネットワーク構造を有するモデルは，画像や自然言語などの非構造データの分析において大きな成果を上げており，学習データの数が多いほど性能が向上する傾向にある．一方で，学習に必要な大量のデータに対するラベル付け作業には膨大なコストがかかるという課題がある．

この課題に対し，なるべく少ないデータ数で性能の高いモデルを構築するための方法として，能動学習が研究されている．能動学習は，モデルの精度向上に寄与する可能性の高いデータを選定し，選定されたデータから順にラベル付けすることで，徐々にラベル付きデータを増やしていく手法である．中でも Yoo ら [1] は，データの選定に損失予測モジュールを用いるアプローチを提案している．このアプローチでは，現段階までにラベル付けされたデータで学習したモデルでは，確からしい予測が困難と推定されたデータを，次にラベル付けするデータとして選定する．そのため，テストデータ全体に対する精度の向上にはあまり寄与しないような，判別が非常に困難な特徴を持つデータも同時に選定してしまい，モデルの性能に悪影響を及ぼす恐れがある．

そこで本研究では，能動学習の学習サイクルにおいて，既にラベル付けされているデータと，ラベル付け候補のデータとの類似度を考慮したラベル付けデータの選定アルゴリズムを提案する．これにより，候補データの分布全体を満遍なく学習することが可能になる．最後に，実データを用いた画像分類タスクに関する実験により，提案手法の有効性を示す．

2. 関連研究

2.1. 能動学習

能動学習とは，なるべく少ないラベル付きデータで高い精度を得ることを目的に，各サイクルにおいて，1) 大量のラベルなしデータの中からモデルの精度向上に寄与する可能性の高いデータを選定，2) 選定されたデータにラベルを付与，3) それまでにラベルが付与された全データを用いてモデルを学習するという手順を繰り返すことで，段階的にモデル学習していく手法である．データの選定方法には，損失予測モジュールを用いたアプローチ [1]，不確実性に注目したアプローチ [2]，データ分布に対するアプローチ [3] などがある．

2.2. 損失予測モジュールを用いたアプローチ

Yoo ら [1] は，現段階で学習されたモデルを用いてまだラベルが付与されていないデータの予測をした際に，損失が大きくなると推定されるデータから優先的にラベル付けをしていく手法を提案している．ここで損失とは，正解ラベルと予測ラベルの差を表すが，ラベル付与候補のデータに対しては, 正解ラベルが付与されていないため損失が計算できない．そこで，ラベルを予測するターゲットモデルとは別に，損失の値を予測するための損失予測モジュールを構築し学習している点が特徴である．

3. 提案手法

3.1. 概要

損失予測モジュールによるアプローチでは，現段階で学習されたモデルを用いて予測した際に，損失が大きくなると予想されるデータに対してラベル付けを行う．しかし，そもそも判別が困難な識別境界付近のデータの予測損失も大きくなると考えられ，この種のデータがラベル付け対象として重点的に選択されてしまう可能性がある．例えば，飛行機と鳥を遠くから撮影した画像を判別することは人間でも極めて困難なタスクである．そのようなデータに対して優先的にラベル付けすることは，テストデータ全体に対する精度の向上には寄与しない可能性が高い．また，偏ったパラメータが推定されてしまい，モデルの性能に悪影響を及ぼす恐れもある．

そこで，学習初期段階においては，それまでに学習されたデータと類似度の低い，離れた特徴を持つデータを選定する方法を考える．これにより，データ全体を満遍なく学習でき，精度の向上につながることが考えられる．本研究では，類似度を測る際の各データの特徴量として，ターゲットモデルの最終層の直前に得られるベクトルを採用する．さらに学習済みデータからの距離を測るための指標としてコサイン類似度を用いた手法を提案する．

3.2. 詳細

提案手法は,（ターゲットモデルにおける最終層の直前から得られる）特徴量空間上における広範囲のデータを満遍なく学習するために，コサイン類似度を用いてデータを選定する．アルゴリズムの詳細を以下に示す．

最初に，ラベルなしデータセット $\mathcal{U}$ から K 個のデータをランダムに選定し，ラベル付けをすることで，ラベル付きデータセット $\mathcal{L}$ を構築する．そして，$\mathcal{L}$ から初期のターゲットモデルを学習する．以降，Step1 から Step5 を終了条件を満たすまで繰り返す．なお，ラベルなしデータセットとラベル付きデータセットのサイズをそれぞれ $N_{\mathcal{U}}$,$N_{\mathcal{L}}$ とする．

Step1： ターゲットモデルを用いて，$\mathcal{U}$,$\mathcal{L}$ 内の全データに対する特徴量 $\mathcal{U}_{\boldsymbol{x}} = \{\boldsymbol{x}_1^{\mathcal{U}}, ..., \boldsymbol{x}_i^{\mathcal{U}}, ..., \boldsymbol{x}_{N_{\mathcal{U}}}^{\mathcal{U}}\}$,$\mathcal{L}_{\boldsymbol{x}} = \{\boldsymbol{x}_1^{\mathcal{L}}, ..., \boldsymbol{x}_j^{\mathcal{L}}, ..., \boldsymbol{x}_{N_{\mathcal{L}}}^{\mathcal{L}}\}$ を得る

Step2：$\mathcal{U}_{\boldsymbol{x}}$ の各データに関する特徴量 $\boldsymbol{x}_i^{\mathcal{U}}$ に対し，$\mathcal{L}_{\boldsymbol{x}}$ 内の全ての $\boldsymbol{x}_j^{\mathcal{L}}$ との類似度を計算し，$\boldsymbol{x}_i^{\mathcal{U}}$ と最も類似するラベル付きデータからの最大類似度 $d_i^* = \max_{\boldsymbol{x}_j^{\mathcal{L}} \in \mathcal{L}_{\boldsymbol{x}}} \cos(\boldsymbol{x}_i^{\mathcal{U}}, \boldsymbol{x}_j^{\mathcal{L}})$ を求める

Step3：$\mathcal{U}$ に含まれる全データに関する最大類似度 $\{d_1^*, \ldots, d_i^*, \ldots, d_{N_{\mathcal{U}}}^*\}$ を比較し，値が小さい方から K 個のデータに対してラベル付けを行う

Step4：ラベル付けしたデータを $\mathcal{U}$ から除外し，$\mathcal{L}$ に加えた上で，$\mathcal{L}$ を用いてターゲットモデルを再学習する

Step5：終了条件を満たしていなければ Step1 へ戻る

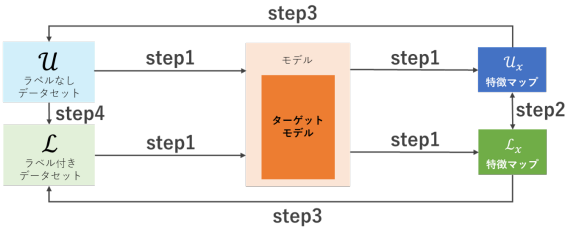

図 1: 提案手法のイメージ

4. 評価実験

4.1. 実験条件

提案手法の有用性を検証するために，CNN を用いた画像分類タスクに関する評価実験を行う．なお，基本的に実験条件は先行研究 [1] に準じ，データセットには CIFER-10 を用いる．CIFER-10 は画像サイズ 32 × 32 × 3，10 のカテゴリからなる物体カラー画像のデータセットである．計 6 万枚の画像データのうち，5 万枚を学習データ，1 万枚をテストデータに用いる．ターゲットモデルには，オリジナルの ResNet-18 から最初の畳み込み層とプーリング層を変更したものを用いる．また，評価指標には正解率を用いる．

能動学習の各サイクルにおいて，ミニバッチサイズ 128，初期学習率 0.1，エポック数 150 として，ターゲットモデルを学習する．能動学習の各サイクルでラベル付けするデータ数 K は 10 とし，計 100 サイクルの学習を行った．比較手法には損失予測アプローチを用いる．

4.2. 実験結果

各手法に対して同じ初期ラベル付きデータセットを用いて，3 回ずつ実験を行った際の正解率の平均値を図 2 に示す．

図 2 より，提案手法は部分的には従来手法よりも精度が下回るものの，ラベル付きデータが 1,000 個の段階 (100 サイクル目) では，従来手法に比べて提案手法が良い結果を示すことがわかる．また，1 から 25 サイクル付近までと 30 から 50 サイクル付近（図 2 の水色部分）においては，特に従来手法よりも高い性能を示した．

5. 考察

提案手法は，従来手法より多くのサイクルで良い精度を得ることができた．特に初期段階においては，損失予測によるアプローチのように間違えやすいデータを重点的に学習するよりも，分布全体に対して広く満遍なく学習することが，精度向上により貢献したと考えられる．また，従来手法ではターゲットモデルだけでなく損失予測モジュールも学習する必要があるのに対し，提案手法ではターゲットモデルのみを学習

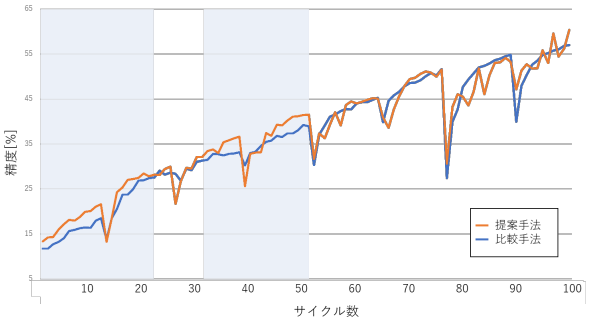

図 2: 各サイクルにおける正解率の推移（3 回の平均値）

すれば良いため，推定するパラメータ数が減少し，早い段階から精度が良くなったと考えられる．

ただし提案手法では学習サイクルによっては精度が下がることもあり，特に 80,90 サイクル付近では従来手法に大きく劣るなど，不安定な結果となった．これは，各サイクルで選定するデータ同士の類似度が高い可能性があることが，原因の 1 つであると考えられる．より効果的な能動学習を実現するためには，1) そのサイクルまでの最も精度の高いモデルを採用する方法，2) 各サイクルで選定するデータ同士の類似度を考慮する方法，3) 初期段階では提案手法を用いた上で，ある程度学習が進んだ段階から損失予測モジュールを用いたアプローチに切り替える方法などが有効であると考える．

また，2 つの手法において，約 13 サイクル周期で精度が大きく下がっているが，これはバッチサイズの設定に原因がある．各サイクルで学習データに追加されるデータが 10 であるのに対し，バッチサイズは 128 であるため，13 サイクル周期でバッチサイズに大きく届かないミニバッチが発生し，そこに対して過学習すると考えられるためである．これは本研究の実験が対象としている，ラベル付けするデータの数を極限に小さくした状況特有の問題であり，文献 [1] を含むほとんどの従来研究では，追加データ数を 10 ほど小さくする状況は考えていないため発生しない．本研究の成果により，今後この未解決の問題に対応するための研究が発展することが期待される．

6. まとめと今後の課題

本研究では能動学習に，既にラベル付けされているデータと，ラベル付け候補のデータとの類似度を考慮してラベル付け対象のデータを選定する，新たな能動学習の方法を提案した．CNN を用いた画像分類タスクに関する評価実験の結果から，提案手法の有効性を示した．今後の課題としては，周期的な精度の低下を解消するためのバッチサイズの変更方法の検討，学習時間の短縮方法の検討などが挙げられる．

参考文献

[1] D.Yoo and I.S.Kweon. Learning loss for active learning. *In Proceedings of CVPR*, pp. 93–102, 2019.

[2] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. *In Proceedings of ACM SIGIR*, pp. 3–12, 1994.

[3] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. *In ICLR*, 2018.