

じん肺多クラス分類におけるデータ数不均衡問題を考慮した学習

程野 仁登 【 知能情報学研究室 】

1 はじめに

機械学習において、不均衡データへの対処はモデルの汎化性能を左右する重要な課題である。本研究で用いる4ステージ多クラスじん肺データにおいても顕著な不均衡がある。不均衡データによる学習では、少数クラスに対するモデルの識別精度の低下をもたらすことがあるため、学習前にデータセットのバランスを調整するリサンプリング手法をはじめとした様々な研究が行われている。本研究では、クラス分布が不均衡なじん肺データセットに対し、不均衡データに対して有効な損失関数である FocalLoss を適用するとともに、訓練データを人工的に生成する手法である SMOTE を深層学習向けに拡張した DeepSMOTE を用いる。さらに、学習時にデータ拡張を行い、これらの手法およびデータ拡張の有無が分類性能に与える影響を明らかにする。

2 実験

2.1 データセット

本実験で用いたデータセットは、高知大学医学部 (KM) の胸部 X 線画像 95 枚からなる (肺野領域抽出済)。各クラスの内訳は、NF (No Finding) が 4 枚、じん肺の重症度に基づくステージ 1 が 34 枚、ステージ 2 が 38 枚、ステージ 3 が 19 枚である。本データセットは、ホールドアウト法および層化 3 分割交差検証法を用いて訓練データ、検証データ、テストデータに分割する。

2.2 実験内容

本実験では、VGG16 を転移学習させたじん肺 4 クラス分類モデルにおいて、損失関数に交差エントロピー誤差を用いる場合をベースラインとする。本実験において、不均衡データに対する手法を用いた学習の流れを図 1 に示す。FocalLoss[1] では、分類が容易なサンプルの重みを下げることで、多数クラスに偏った学習を抑制し、クラス不均衡に対処する。また、DeepSMOTE[2] では、少数クラスの訓練データを生成することで、クラ

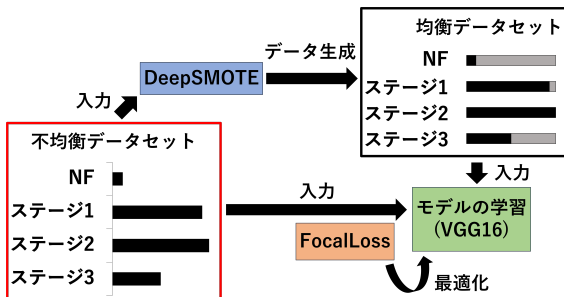


図 1 不均衡データに対する手法を用いた学習の流れ

不均衡修正手法	データ拡張	マクロ平均 F1 スコア
Baseline (修正なし)	なし	0.190
	あり	0.228
FocalLoss (FL)	なし	0.424
	あり	0.458
DeepSMOTE (DS)	なし	0.279
	あり	0.352
FL + DS	なし	0.413
	あり	0.346

ス分布を均衡化し、分類モデルの学習を行う。さらに、学習時にデータ拡張を行い、その有無による分類性能の差異を比較する。なお、FocalLoss および DeepSMOTE を用いる際のハイパーパラメータは、検証データに基づいて最適化する。性能評価には、各クラスの F1 スコアの平均であるマクロ平均 F1 スコアを用いる。

3 結果・考察

各手法およびデータ拡張の有無による実験結果を表 1 に示す。ベースラインと比較し、FocalLoss を適用したモデルは、データ拡張の有無にかかわらず大きな性能向上を示した。一方、DeepSMOTE のみを適用した場合には、FocalLoss を用いた場合ほどの改善は見られなかった。これは、DeepSMOTE の学習に用いられるデータ数が十分でなく、有効なデータ生成が十分に行われなかった可能性によるものと考えられる。また、FocalLoss と DeepSMOTE を併用した場合には、FocalLoss のみの場合と比較し性能向上は確認されず、データ拡張を併用した条件では性能の低下が見られた。このことから、不均衡データに対する複数の手法の単純な併用は必ずしも有効ではないことが示唆される。

4 まとめ

本研究では、じん肺多クラス分類における不均衡データ問題に対し、FocalLoss および DeepSMOTE を適用し、データ拡張の有無を含めて有効性を比較した。その結果、FocalLoss とデータ拡張を併用した場合に最も高い分類性能が得られた。一方、DeepSMOTE は本実験条件下では十分な性能向上を示さず、データセットの規模や特性に依存する可能性が示唆された。

参考文献

- [1] T.-Y. Lin et al., ICCV, 2017.
- [2] D. Dablain et al., 2023.