Vol.2019-MPS-126 No.9
2019/12/11

ディープラーニングを用いたデータモデリングにおける汎化性能評価とその改善法に関する検討

Evaluation of generalization in data modeling with deep learning and its tuning methods

矢野 浩史[1], 新川 裕也[2],石井 一夫[2]
久留米大学大学院医学研究科[1], 久留米大学バイオ統計センター[2]
Hiroshi Yano[1], Yuya Shinkawa and Kazuo Ishii[2]
Graduate School of Medicine, Kurume University[1],
Biostatistics Center , Kurume University[2]

はじめに

医療分野におけるデータモデリングでは、推定した数理モデルに対する訓練データの当てはまりのよさを指標に最適化モデルが選択されることが多い。これらの最適化モデルには過学習が起こっている可能性が高いが、その汎化性能に関して検討されることはあまりない。

講演者らの研究グループにおける先行研究として、健診データによる脳白質病変予測に関するものがあり[1)]、様々なデータ分析手法で訓練データに対する最適化がなされた。当該研究では、最適化モデルとしてロジスティック回帰モデルを選択した。しかし、このモデルは過学習については検討されていないためモデルの最適化が完全になされているとは言い難い。そこで、TensorFlow で実装したロジスティック回帰モデルを用いて、損失誤差を指標に過学習を検討し、ドロップアウトや L１正則化などの過学習を減弱させるための手法について検討を行ったので、ここに紹介する。

方法

(1) 使用したデータセット

先行研究[1)]の 1914 人分（男性 988 人、女性 916 人）の健診データのデータセットを用いた。説明変数には、連続型変数として、年齢, プラークスコア(PS), LDL コレステロール(LDL), HbA1c, 収縮時血圧(SBP)を、カテゴリカル変数として、性別、メタボ判定、降圧剤投与の有無、インシュリン投与の有無、コレステロールを下げる薬剤の投与の有無、飲酒習慣の有無を用いた。目的変数には、脳白質病変の有無を用いた。連続型変数は正規化を行い、カテゴリカル変数はダミー変数化して以後の解析を行った。

(2) モデルの構築

TensorFlow および Keras によりロジスティック回帰モデルを構築した。損失関数として交差エントロピー誤差を用い、勾配降下法におけるオプティマイザとして Adam (Adaptive moment estimation)を用いた。機械学習モデルの評価には、Accuracy を用いた。モデルの活性化関数として、中間層には ReLU (Rectified Linear Unit)を、出力層にはシグモイド関数を用いた。

(3) モデルの検討項目

損失誤差が少なく、より過学習の少ない汎化性能の優れたシンプルなモデルを選択するため、中間層のノード数の最適化を行った。さらに、過学習の少ないモデルへのチ

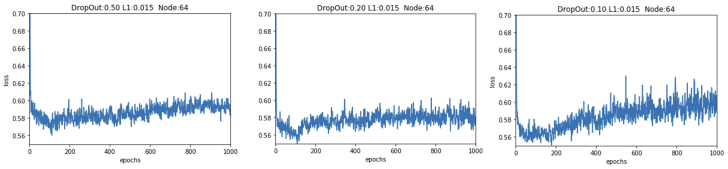
表１　各分析手法の性能比較

	AUC	カットオフ値	正診率	誤診率	感度	特異度	PPV	NPV
LogReg(DL)	0.813	0.433	77.1%	22.9%	86.9%	64.5%	75.9%	79.3%
LogReg	0.799	0.566	71.1%	28.9%	64.4%	79.4%	79.1%	64.8%
NB	0.776	0.382	72.0%	28.0%	76.5%	66.6%	73.8%	69.8%
SVM	0.787	0.679	70.7%	29.3%	64.5%	78.7%	48.8%	28.0%
RF	0.79	0.428	71.7%	28.3%	83.1%	58.0%	39.8%	30.6%

ューニングのために、ドロップアウトおよび、L１とL２正則化の検討を行った。

(4) モデルの識別性能評価

最終的に得られたチューニングモデル(LogReg (DL))の識別性能を評価するため、正診率、感度、特異度、陽性的中率(PPV)、陰性的中率(NPV)、ROC 曲線(ROC 曲線下の面積(AUC)を含む)などを、Shinkawa らの論文[1)] のロジスティック回帰モデル(LogReg)、ナイーブベイズ(NB)、サポートベクトルマシン(SVM)、ランダムフォレスト(RF)と比較した。



左：DropOut=0.5　　中：DropOut=0.2　　右：DropOut＝0.1

図１　機械学習モデルにおける DropOut 率の検討

結果

ディープラーニングによるロジスティック回帰モデルについて node 数, DropOut 率, L1, L2 正則化などのモデルの設定値の最適化を行った。その結果、node 数=64, DropOut 率=0.2, L1=0.015 のとき、交差エントロピー誤差=0.525, 正診率=0.813 となり最良の結果を得た。このうち、DropOut 率=0.2 を選択したときの検討例を図１に示した。図１のグラフは縦軸に損失誤差を、横軸に epochs (試行回数)を取っている。真ん中の DropOut 率=0.2 のとき、損失誤差の極小値が最も小さく、過学習も DropOut 率=0.1 の時より少ないことが確認できる。

各手法の識別性能の比較を表１に示す。ディープラーニングでチューニングしたモデル（LogReg(DL)）は、AUC が 0.813、正診率が 77.1%、感度が 86.9%、NPV が 79.3%となり検討した各モデルの中で最大であった。特異度、PPV は、LogReg よりやや劣るが、総合的にみて脳白質病変をより高精度に捉えることができるモデルであった。

考察

ディープラーニングにより過学習のより少ない汎化性能の高いモデルを作成する方法を確立した。得られた最適化モデルは、従来のモデルより識別性能の改善が見られ、検討した中では、最も高い正診率を示した。その結果、汎化性能を改善することで、識別性能が改善されることが示された。

引用文献

1) Shinkawa Y *et al*. Mathematical modeling for the prediction of cerebral white matter lesions based on clinical examination data *PLoS ONE* 14(4): e0215142 (2019)