
ポスター発表

[PB] ポスター B

2020年6月5日(金) 09:00 ~ 18:30 ポスター会場(1) (e-poster)

[PB-16] 医療データの特徴を考慮した多施設間 Federated Learningの提案 Proposal for Federated Learning Based on Medical Data Features

*馬 雲蔚¹、岡本 和也^{1,2}、杉山 治²、山本 豪志朗²、佐々木 博史²、南部 雅幸²、黒田 知宏^{1,2} (1. 京都大学大学院情報学研究科、2. 京都大学医学部附属病院)

*Yunwei Ma¹, Kazuya Okamoto^{1,2}, Osamu Sugiyama², Goshiro Yamamoto², Hiroshi Sasaki², Masayuki Nambu², Tomohiro Kuroda^{1,2} (1. Graduate School of Informatics, Kyoto University, 2. Kyoto University Hospital)

医療データの特徴を考慮した 多施設間 Federated Learning の提案

MA YUNWEI^{*1}, 岡本和也^{*1,2}, 杉山治^{*2}, 山本豪志朗^{*2},
佐々木博史^{*2}, 南部雅幸^{*2}, 黒田知宏^{*1,2}

^{*1} 京都大学大学院情報学研究科, ^{*2} 京都大学医学部附属病院

Proposal for Federated Learning Based on Medical Data Features

Yunwei Ma^{*1}, Kazuya Okamoto^{*1,2}, Osamu Sugiyama^{*2}, Goshiro Yamamoto^{*2},
Hiroshi Sasaki^{*2}, Masayuki Nambu^{*2}, Tomohiro Kuroda^{*1,2}

^{*1} Graduate School of Informatics, Kyoto University

^{*2} Kyoto University Hospital

抄録: 医療分野で深層学習の応用が試みられているが、医療分野で深層学習を行う際の問題点として、学習データとする医療データが各施設に散在しており、患者のプライバシーに配慮しつつ大規模に学習データを集めることが難しいという点が挙げられる。この問題に対する解決策として、Federated Learning という手法を利用することが考えられる。Federated Learning は学習データを共有せず、学習モデルの重みのみを共有して学習を行う。しかし、現在 Federated Learning に関する研究の大半は学習の高速化を目的にしており、医療分野への応用について提言はされているものの、医療データの特徴を考慮した手法の提案はなされていない。本研究では、まず糖尿病眼底写真データを使用し Federated Learning の有効性を検証し、医療データの特徴を考慮した Federated Learning の学習手法の考察を行った。

キーワード 深層学習、Federated Learning、医療データ、プライバシー

1. はじめに

一般画像認識において深層学習が高い精度を達成した 2012 年以来、医療分野でも深層学習の応用が進んでいる。深層学習の精度向上には学習データの量が重要であるが[1]、医療データは各施設が独自に保有しており、医療データに個人情報が含まれているため、大量の学習データを収集することは難しい。

この問題に対応するため、Federated Learning[2, 3]を利用することが考えられる。Federated Learning はトレーニングデータをローカルに保持するクライアントが多数参加して共有グローバルモデルをトレーニングする、機械学習に対するアプローチである。Federated Learning では多施設の医療データを収集することなく、学習モデルの重みのみを共有することで学習を行えるため、患者のプライバシーに配慮しながら学習データを相互に利用することが可能である。しかし、Federated Learning に関する研究は、データのプライバシー保護や制限になっている通信コストなどの学習プロセスの改善に焦点を当てており、[3, 4,

5]、医療データの特徴を考慮した手法の提案はなされていない。

本研究の目的は、実際に医療データが複数施設に分散している環境を想定し、分散した医療データに Federated Learning を適用することで、Federated Learning が医療データに対しても有効であるかを検証する。

2. 方法

2.1 実験に用いるデータ

本研究ではインターネット上で公開されている糖尿病眼底写真データ[6]を使用する。糖尿病眼底写真データは糖尿病性網膜症を検出するための画像である。評価尺度は「No DR」「Mild」「Moderate」「Sever」「Proliferative DR」の 5 段階に分かれていて、左目と右目のデータが混在している。今回は「No DR」を負例、「Sever」と「Proliferative DR」を正例とし、2 値ラベル(0,1)に変換し、全 1,903 のデータを使用する。

2.2 想定した環境

本研究では、学習条件を 3 つに分けて行なった。条件 1 では全データをまとめて学習する。条件 2

ではデータをシャッフルし、データ数が多い施設 A とデータ数が少ない施設 B という環境を想定し、データ比率を 4:6 に分けて別々に学習する。条件 3 では条件 2 で分けたデータに Federated Learning を適用する。仮説として、条件 3 での正解率は条件 1 での正解率とほとんど変わらず、なおかつ条件 2 での正解率より良くなると考えた。

3. 結果

実験では 6 回の試行を行い、正解率の平均をとった(Table.1)。糖尿病眼底写真データによる実験は上述の仮説通りの結果を示した。

また、条件3における混同行列を Figure.1 に示す。混同行列を見ると、負例と正例の正解数に偏りがあることが示された。

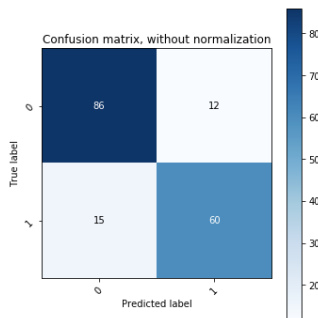


Figure.1 条件 3 における混同行列

Table.1 実験正解率

	条件 1	条件 2 (施設 A)	条件 2 (施設 B)	条件 3
糖尿病眼底写真データ	81.98%	80.35%	77.46%	82.66%

4. 考察

1. 正解率について

Table.1 から、糖尿病眼底写真データに関する Federated Learning の有効性の検証においては、仮説通りの順番で条件の正解率が推移したが、Figure.1 の混同行列を見ると、学習がデータ量の多いラベル「0」に偏っていることがわかる。条件 1 での実験では、全体の正解率が上がるにつれ各ラベルの識別もより適切になっているが、条件 3 では、ラベルごとの識別結果に偏りがあることがあった。この原因としては、現状の Federated Learning が通信コストの削減に焦点を当てており、各施設が保持する医療データのデータ量の不均衡などの特有の性質を考慮していないことが考えられる。

2. 医療データ特有の性質

実際の医療施設の特性を考えたとき、施設の規模・専門性によって、同じ病気でも症例数は大きく異なると考えられるため、施設間のデータ数のばらつきは大きい。また、特定の分野に特化した専門医療施設では、同じ施設内のデータのバイアスが大きい可能性がある。このような施設では、希少データが存在する可能性が高く、このようなデータのバイアスについても学習において考慮する必要がある。また、逆に特定の分野に特化した医療施設では、似たような特徴を持つデータが集まることも考えられる。このように、医療施設における Federated Learning では、通信コストだけでなく、施設間のデータの不均衡、及び、データ数の偏りを考慮した学習手法を考案することで、より良い精度の識別器が得られると考えられる。

5. 結語

多施設に分散されている医療データを学習するには Federated Learning は必要だが、現存の手法では実環境での医療データに十分でないと考えられる。今後、医療データを保持する施設がもつ「データの不均衡」や「データ数の偏り」などの問題に着目し、さらに適する手法を提案する必要がある。

参考文献

- [1] Uchida, S., et al.: A further step to perfect accuracy by training CNN with larger data, Proc. ICFHR 2016, 405-410, 2016.
- [2] YANG, Qiang, et al.: Federated machine learning: Concept and applications. ACM TIST, 10.2: 1-19, 2019.
- [3] ABADI, Martin, et al.: Deep learning with differential privacy.: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308-318, 2016.
- [4] Tao, Z., et al.: eSGD: Communication efficient distributed deep learning on the edge, Proc. HotEdge 18, 2018.
- [5] Konečný, J., et al.: Federated learning: Strategies for improving communication efficiency, arXiv, arXiv:1610.05492, 2016.
- [6] Kaggle: Diabetic Retinopathy Detection, <https://www.kaggle.com/c/diabetic-retinopathy-detection>, (accessed 2020-02-25)