The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

訓練データの改ざんを考慮した機械学習手法

The Method of Machine Learning Considering Tamper of Training Data

柳町 天使　　石田 好輝
Tenshi Yanagimachi　Yoshiteru Ishida

豊橋技術科学大学
Toyohashi University of Technology

Big data is increasingly used as training data for machine learning. However, large-scale data such as big data is not always appropriate as training data at all times. Particularly when collecting data from Web services such as SNS, unspecified number of people using the service can indirectly tamper with data. In this research, we propose a learning method and verify its effectiveness so as to obtain a learning result close to the case without tampering even in an environment in which part of the training data has been tampered with. In this method, learning is divided into two stages, and the reliability of the training data is estimated using the first stage learning result, thereby assisting the exclusion of tampered data by human beings.

1. はじめに

近年，あらゆる分野で AI（artificial intelligence）の活用が広がっている．AI の実現方法には様々なものがあるが，主に用いられているのはニューラルネットワークを用いて機械学習を行う形式のものである．機械学習では一般的に，訓練データが多いほど学習後の性能が高くなる傾向にある．そのため，ビッグデータのような大規模なデータを訓練データとして用いることも多い [Wu 16]．

ビッグデータは多くの訓練データを用意できるという点で機械学習と相性が良い．しかし，そのような大規模データを用いることで発生する問題もある．特に SNS のような Web サービスからデータを集める場合，サービスを利用する不特定多数の人間が間接的にデータを改ざん可能になる．この性質を悪用された事例として，Microsoft が Twitter 上で公開していたチャットボット Tay が不適切な発言を繰り返すようになったというものがある [Mizoroki 16]．また Google 翻訳でも同様の問題により，本来の訳語とは全く関係のない語に翻訳される不具合が起きている [インターネットコム編集部 17]．このような問題に対して，現在はフィルタリングによる対策が行われている [Lee 16]．しかし，人間の作るフィルターでは望まない入力や出力を全て除外することは難しい．また，AI の活用が広がる中で，今後は自動車の自動運転などミスが許されないタスクにも AI が用いられることが考えられる．よって訓練データの改ざんが行われることを前提とした機械学習手法の確立が求められる．

ラベルにノイズが含まれる環境での機械学習については多くの先行研究が存在する．しかし，想定する状況において先行研究の手法を用いる場合いくつかの解決すべき問題がある．第一に適切な手法がタスクの種類に依存すること．第二に訓練データとは別にノイズの定義が必要であること．第三にノイズの定義が不完全なとき，ノイズの除去も不完全になってしまうこと．そこで本研究では次の二点を満たす機械学習手法の開発を目標とした．一点はタスクに依存せず用いることのできる手法であること．そしてもう一点は，同様の処理を繰り返すことで最終的にノイズのない環境と同程度の学習結果を得られることである．上記の目的を達成するため，1 つの手法の提案と有用性の検証を行った．この手法は，半教師あり学習のように学習を 2 段階に分け，1 段階目の学習結果を用いて訓練データの信頼度を推定することで，改ざんされたデータの除外を試みるものである．

連絡先: 柳町 天使，豊橋技術科学大学 大学院工学研究科

2. 関連研究

2.1 ラベルノイズの存在下における分類

ニューラルネットワークを用いた機械学習の分野において，ラベルノイズの存在する環境下における学習手法の提案は数多く行われている．Frénay and Verleysen[Frénay 14] はこれらの手法を特徴によって次の 3 種類に分類している．

- noise-robust methods（ノイズロバストな手法）
- data cleansing methods（データ浄化による手法）
- noise-tolerant methods（ノイズ耐性のある手法）

3 種類の手法は，ノイズの特性を仮定することで対策を行っている．ノイズロバストな手法では，ノイズの影響は過学習に陥った場合を除いて無視できるものと仮定している．そして過学習自体を回避することでノイズの影響の排除を試みている．データ浄化による手法では，ノイズの付加されたラベルと正常なラベルを判別するためノイズの定義が得られる環境を仮定している．ノイズ耐性のある手法では，ネットワークの構造を工夫することでノイズに対して学習結果が鋭敏に反応しないようにしている．しかし結果的に正常なラベルに対する学習も鈍くなる傾向がある．

また，Jindal, Nokleby, and Chen[Jindal 16] は巨大かつ信頼性の低いデータセットを訓練データとする場合の学習手法を提案している．この手法ではノイズモデルを学習することで正しい分類を行うことを試みている．

これら先行の手法に共通する特徴は，訓練データにどのようなノイズが付加されるのか予想可能な環境での学習であるという点である．本研究において想定する，不特定多数の人間が意図的に改ざんを行う環境ではどのようなノイズが訓練データに付加されるかを予想することが困難である．そこで本研究では，人間がラベルの正常な訓練データをある程度用意することで改ざんされたデータを判別できるニューラルネットワークを学習させる．

2.2 Ugly Duckling Theorem

Watanabe, Knowing and guessing[Watanabe 69] は Ugly duckling theorem（みにくいアヒルの子の定理）として分類問題に関する定理を示している．この定理は複数の対象において全ての特徴量を同等の重みで扱う限り，対象の分類は行えないというものである．機械学習では特徴選択や特徴抽出の基準，各特徴量の適切な重み等を訓練データから学習することで取得する．そのため正常な環境では Ugly duckling theorem で示されるような分類不可能な状態には陥らない．しかし，今回想定するような訓練データに初めから誤ったデータが含まれている環境では適切な基準や適切な重みが得られない．よって分類不可能な状態を回避するためには改ざんされた訓練データとは別の，信頼できる訓練データが必要となる．

2.3 半教師あり学習

本研究では改ざんを考慮した学習手法として，半教師あり学習を参考とした手法を提案する．半教師あり学習はラベル付けされたデータに加えて，ラベルの付加されていないデータを訓練データとして用いることで学習結果の性能を高めることを目的とした手法である [Nigam 98]．このラベル付けされたデータとラベルの付加されていないデータという区別を，提案手法ではラベルの信頼できるデータと信頼できないデータという区別に置き換えて考える．提案手法では信頼できるデータで学習した結果を用いて信頼できないデータのラベルの正しさを推測することで，改ざんの影響を少なく抑える．

3. 手書き数字認識タスク

3.1 タスク概要

本稿では提案手法の有効性の検証のため，タスクとして手書き数字画像の認識タスクを用いる．このタスクは，人間が手書きした「0」～「9」のアラビア数字を認識し適切に分類するというものである．以降に学習のため用いるデータセット，ネットワーク構造，そして訓練データに対するラベルノイズの付加方法について詳細を示す．

3.2 MNIST データセット

訓練データおよびテストデータには MNIST データセット [LeCun 98] を用いる．MNIST データセットは訓練データ 60,000 枚，テストデータ 10,000 枚からなる手書き数字データで，1 枚は高さと幅がそれぞれ 28 pixel のグレースケール画像データである．MNIST データセットに含まれる手書き数字画像の例を図 1 に示す．

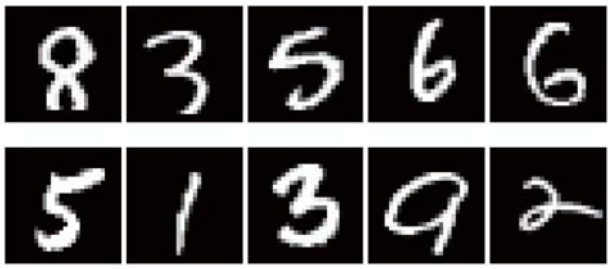

図 1: MNIST データセットに含まれる手書き数字画像の例

3.3 ネットワーク構造

学習用のニューラルネットワークには 1 層の中間層に softmax 層を合わせたものを使用する．ネットワークの構造を図 2 に示す．また，ネットワークの出力値の定義を式 1 に表す．

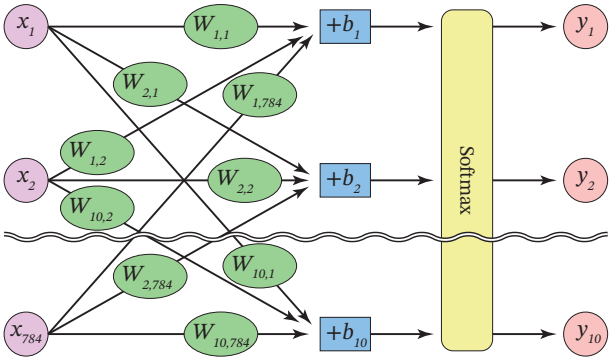

図 2: ニューラルネットワーク構造

$$y_i = softmax(\sum_j W_{i,j}x_j + b_i) \quad (1)$$

y... 出力値

x... 入力値（画像 1 pixel の色情報，0～255）

W... 結合強度

b... バイアス

i... ラベルインデックス（1～10）

j... 画素インデックス（1～784）

3.4 ノイズの付加

本稿において訓練データへのノイズの付加は次のように行う．まず，対象となるデータ群から無作為に一定枚数選出する．次に，選出したデータに対応付けられたラベルを正常なものから 1 つ値を増加させる．つまり正常な「1」のラベルは必ず「2」に書き換えられ，「9」のラベルは「0」に書き換えられる．ランダムな値に書き換えずラベルの変化に法則性を持たせることは，悪意を持った人間の意図的な改ざんを再現することを意図している．

4. 提案手法

初めに提案手法に用いる各用語の定義について示す．次に，提案手法による学習方法について説明を行う．

4.1 用語定義

4.1.1 作成者

タスクを解決するためのニューラルネットワークを作成する人間．訓練データに関してラベルの正しさを判断する基準を持つ．

4.1.2 信頼データ・非信頼データ

改ざんによるラベル誤りが存在する可能性のあるデータを非信頼データとする．そのため，学習の初期状態では全ての訓練データが非信頼データとなる．また，作成者によってラベルが正しいと判断された，もしくは正しいラベルが付け直されたデータを信頼データとする．

4.1.3 振り分けニューラルネットワーク（振り分け NN）

信頼データのみを訓練データとして学習を行ったニューラルネットワーク．非信頼データを後述の仮信頼データ・仮非信頼データ・曖昧データに分類するために用いる．ネットワークの構造は後述の最終ニューラルネットワークと同一のものとする．

4.1.4 仮信頼データ・仮非信頼データ・曖昧データ

非信頼データを振り分けニューラルネットワークの出力値を基準に分類したもの．仮信頼データは十分ラベルが正しいと予想されるもの．仮非信頼データはラベルが誤りであると予想されるもの．曖昧データはその段階で判断の付けられないものとして分類する．それぞれに振り分ける具体的な基準値はタスクごとに設定する．

4.1.5 最終ニューラルネットワーク（最終 NN）

信頼データと仮信頼データを合わせたものを訓練データとして学習を行ったニューラルネットワーク．実際にタスクを処理するために用いる．ネットワークのノード数，階層数といった構造は処理するタスクにより適切なものを用いるが，提案手法は最終ニューラルネットワークがどのような構造であっても利用できることを目標としている．

4.2 学習方法

提案手法による学習方法の概要を次のリストに示す．提案手法ではまず，振り分けニューラルネットワークの学習を行うために信頼データの確保を行う．次に振り分けニューラルネットワークによる非信頼データの分類を行う．振り分けニューラルネットワークによる非信頼データの分類は，ニューラルネットワークの出力値を基準として行う．本稿における以降の実験では，振り分けニューラルネットワークの出力が 0.5 を超過するものを仮信頼データに，0.1 未満のものを仮非信頼データに，その他のものを曖昧データに振り分けている．その後，仮非信頼データと曖昧データから一定数のデータを選択，作成者が正しいラベルに修正し信頼データに追加する．訓練データ全体からではなく仮非信頼データ，曖昧データから修正を行うことで，振り分けニューラルネットワークの学習を効率的に進める．提案手法において重要なことは，信頼データが作成者によってラベルの正しさを保証されたデータのみで構成されていること，そして振り分けニューラルネットワークが信頼データのみを元に訓練されることである．これにより振り分けニューラルネットワークは作成者の示す基準によって対象となるデータの正しいラベルを学習する．最終的に振り分けニューラルネットワークが正常だと判断した仮信頼データと信頼データを合わせて訓練データに用いることで，最終ニューラルネットワークは学習を行う．

1. 訓練データを信頼データと非信頼データに分割，信頼データはラベルを修正
2. 信頼データを元に振り分けニューラルネットワークを学習
3. 振り分けニューラルネットワークを用いて非信頼データを仮信頼データ・仮非信頼データ・曖昧データに分類
4. 仮非信頼データおよび曖昧データから一定数のデータを選択，ラベルを修正し信頼データに追加
5. 残った仮信頼データ・仮非信頼データ・曖昧データを合わせ，新たな非信頼データとして定義
6. （2）～（5）を 1 サイクルとして任意の回数実行
7. 信頼データと仮信頼データを合わせて訓練データとして最終ニューラルネットワークを学習

5. 改ざん率に対する認識精度の検証

5.1 概要と目的

訓練データが改ざんされた環境において提案手法によるラベル修正を行った場合，改ざんのない環境と同程度の学習結果が得られるかを検証する．そのために，一定数の訓練データにノイズを付加し学習を行った際の最終的な認識精度を記録した．

5.2 条件設定

実験条件の概要を表 1 に示す．訓練データの改ざん率を 0～100%の間で変化させる．このときの最終ニューラルネットワークの認識精度を確認する．

表 1: 改ざん率に対する認識精度の検証における条件設定

訓練データ	信頼データ（1,000 枚）＋仮信頼データ
テストデータ	MNIST テストデータ（10,000 枚）
ラベル改ざん率	0～100%（10%刻み）
初期の信頼データ数	10 枚
1 サイクル毎の信頼データ追加数	30 枚

5.3 結果

テスト用データセットを用いて認識精度を確認した結果を図 3 に示す.「提案手法」「ラベルの修正をしない場合」それぞれのグラフは，各条件において 5 回試行した平均を表している．また,「改ざんのない環境での平均認識精度」は訓練データにノイズの付加されていない環境において，用いるニューラルネットワークを愚直に学習させた場合の平均的な認識精度を参考として示したものである．図より，提案の手法では訓練データの 8 割が改ざんされた環境であっても改ざんのない環境と同程度の認識精度を維持できていることを確認した．

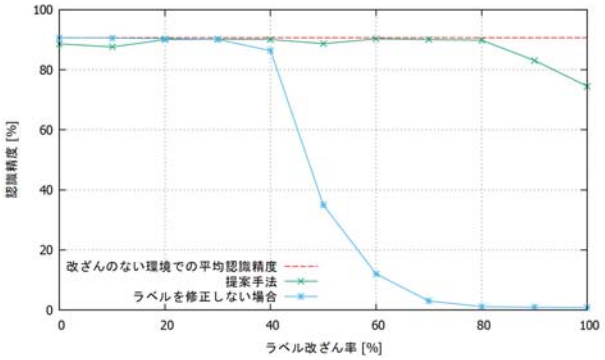

図 3: 無作為なラベル修正と提案手法との認識精度の変化の比較

6. 判断回数に対する認識精度の検証

6.1 概要と目的

提案手法において，作成者によるラベル判断回数を少なく抑えつつ，最終ニューラルネットワークの認識精度を向上できているかを検証する．そのために，作成者によるラベル判断回数を増加させながら，最終ニューラルネットワークの認識精度の変化を記録した．

6.2 条件設定

実験条件の概要を表 2 に示す．本実験では元となる訓練データの 1 割にラベルの誤りを付加している．

6.3 結果

テスト用データセットを用いて認識精度を確認した結果を図 4 に，ラベル修正回数 10～310 回の区間について図 5 示す．それぞれのグラフは「ランダム修正＋仮信頼データの追加なし」

表 2: 判断回数に対する認識精度の検証における条件設定

訓練データ	信頼データ（10〜1,000 枚）＋仮信頼データ
テストデータ	MNIST テストデータ（10,000 枚）
ラベル改ざん率	10%
初期の信頼データ数	10 枚
1 サイクル毎の信頼データ追加数	30 枚

「ランダム修正＋仮信頼データの追加あり」「提案手法による修正＋仮信頼データの追加なし」「提案手法による修正＋仮信頼データの追加あり」の各条件において５回試行した平均を表している．また，ランダム修正とは訓練データ全体から無作為に選んだデータのラベルを修正する方法を表し，仮信頼データの追加なし（あり）とは最終ニューラルネットワークの訓練データとして仮信頼データを信頼データに追加するか否かを表す．図より「提案手法による修正＋仮信頼データの追加あり」の場合が最も早く改ざんのない環境と同程度の認識精度（約９０％）に達していることが分かる．

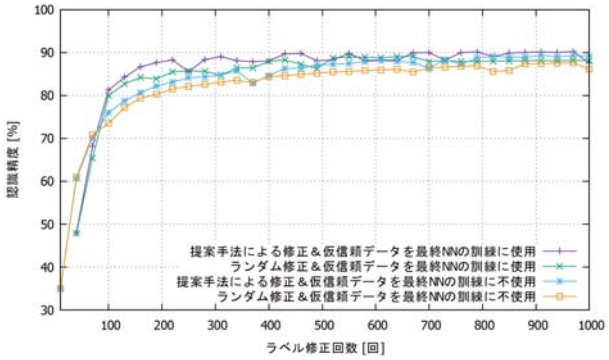

図 4: ラベル判断回数を基準とする認識精度の推移の比較

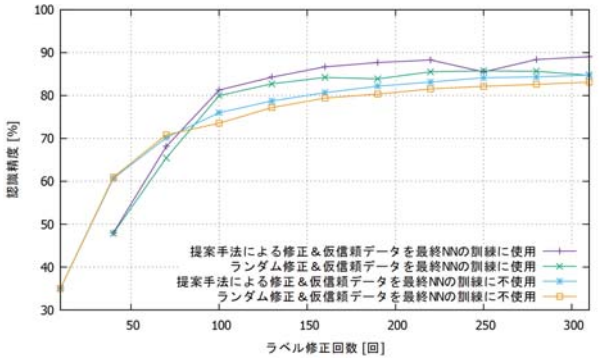

図 5: ラベル判断回数を基準とする認識精度の推移の比較（ラベル修正回数 10〜310 回区間）

7. 考察

改ざん率に対する認識精度の検証の結果より，提案手法によってラベル修正を十分な回数繰り返すことで，訓練データが改ざんされていない環境と同程度の学習結果を得られることが示された．また，判断回数に対する認識精度の検証の結果より，無作為にラベル修正を行う方法と比較して，一定以上の学習結果を得るために必要なラベル修正回数を少なく抑えられていることが示された．以上のことから，提案手法はラベル修正作業に関する作成者の負担を少なく抑えながら，訓練データに対する改ざんのない環境に近い学習結果を得ることが可能であると考える．

8. おわりに

訓練データの改ざんを考慮し，作成者と振り分けニューラルネットワークとでラベルの正誤判断とデータの網羅的検証の役割を分担することで，改ざんのない環境下と同程度の学習結果を得られた．今後は本手法を翻訳等の複雑なタスクに用いて同様の結果が得られるかを確認すること，そして手法をより効率化し作成者の負担をさらに低減できるよう改良を行うことが目標となる．

参考文献

[Frénay 14] Frénay, B.: Verleysen, M.: Classification in the Presence of Label Noise: a Survey, IEEE Transactions on Neural Networks and Learning Systems, 25(5), 845-869, 2014.

[Jindal 16] Jindal, I.: Nokleby, M.: Chen, X.: Learning Deep Networks from Noisy Labels with Dropout Regularization, Proceedings of The 16th IEEE International Conference on Data Mining, 967-972, 2016.

[LeCun 98] LeCun, Y.: Cortes, C.: Burges, C. J.: MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges, http://yann.lecun.com/exdb/mnist/, 1998.

[Lee 16] Lee, P.: Learning from Tay ’s introduction - The Official Microsoft Blog, Microsoft: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/, 2016.

[Mizoroki 16] Mizoroki, S.: Kantrowitz, A.: Microsoft の人工知能は、なぜ虐殺や差別を「支持」するようになったのか……, BuzzFeed Japan: https://www.buzzfeed.com/jp/sakimizoroki/microsofts-ai-chatbot-into-a-neo-nazi-jp, 2016.

[Nigam 98] Nigam, K.: McCallum, A.: Thrun, S.: Mitchell, T.: Learning to Classify Text from Labeled and Unlabeled Documents, Proceedings of The 15th National Conference on Artificial Intelligence, 1998.

[Watanabe 69] Watanabe, S.: Knowing and guessing; a quantitative study of inference and information, Wiley, 1969.

[Wu 16] Wu, X.: Ito, K.: Iida, K.: Tsuboi, K.: Klyen, M.: りんな：女子高生人工知能, 言語処理学会 第 22 回年次大会 発表論文集, 306-309, 2016.

[インターネットコム編集部 17] インターネットコム編集部: Google 翻訳、頼りすぎるのは危険？―「イタズラ」で誤訳の恐れ、なお残る [インターネットコム], internetcom K.K.(Japan): https://internetcom.jp/202511/google-translate-caution, 2017.