DEIM Forum 2023 1A-8-4

クラウドソーシングにおける作業結果のクレンジングによるデータセットの品質向上

三島　惇也†　　鈴木　　優††

† 岐阜大学大学院自然科学技術研究科　〒 501–1193 岐阜県岐阜市柳戸 1-1
†† 岐阜大学工学部電気電子・情報工学科　〒 501–1193 岐阜県岐阜市柳戸 1-1
E-mail: †b4525091@edu.gifu-u.ac.jp, ††ysuzuki@gifu-u.ac.jp

あらまし　教師あり機械学習モデルを学習させる際，教師データの正確性が重要である．しかし，MNIST などのデータセットの正確性は十分であるとはいえない．そのため，ノイズになり得るデータを削除するためにデータクレンジング手法が使用されることがあるが，データ数が減少するという問題点がある．そこで，我々はデータ数を減らすことなくデータセットを作成可能なデータクレンジング手法を提案する．通常は，クラウドソーシングで集めた投票データを集計したデータセットに対してデータクレンジングを行う．それに対し，提案手法ではクラウドソーシングで集めた投票データに対してデータクレンジングを行い，その後に投票データの集計を行う．提案手法を用いることで，データ数を減らさないデータクレンジングが可能であると考えた．本論文では，何もしない場合，通常のデータクレンジングを行った場合，提案手法を用いた場合の 3 パターンの実験を行い，精度の比較を行った．

キーワード　データクレンジング，機械学習，ニューラルネットワーク，クラウドソーシング，外れ値検出

1　はじめに

クラウドソーシングは，教師あり機械学習に必要な教師データを作成するために用いられる．クラウドソーシングでは作業依頼者がインターネット上に依頼を出し，インターネット上にいる作業者が依頼を受けて作業を行う．データセットの作成方法は様々だが，我々が想定しているデータセット作成方法を以下に示す．まず，作業依頼者は学習に使用したいデータを集める．次に，作業依頼者はそれらのデータに付与すべき教師ラベルの投票を作業者に依頼し，複数人の作業者から投票結果を得る．そして，作業依頼者は得られた投票結果をもとに多数決を取り，付与する教師ラベルを決定する．以上の手順でラベル付けされたデータセットを作成する．

上記のデータセット作成方法は，多数派の意見が正しいと仮定したうえで採用されていると考えられるが，実際にはそうではないことがある．なぜなら，質の悪い作業者が一定数存在するからである．例えば，文章を書いた人の感情がポジティブとネガティブのどちらに属するかという質問に対してすべてにポジティブと答える作業者がいたり，内容を確認しないまま回答し，ランダムにラベルを付ける作業者がいたりする．上記のような質の悪い作業者の投票数が正常な作業者の投票数よりも多い場合，質の悪い作業者が投票したラベルが教師ラベルに採用されてしまう可能性があると考えられる．このことから，我々が想定しているデータセット作成方法は正確なラベル付けが行えているとはいえないと考えた．そして我々は，正確なラベル付けが行えていない原因が多数決による教師ラベルの決定にあると考えた．

また，データセットの正確性が不十分であるため，学習に使用するデータセットは教師ラベルにノイズがあることを前提として使用しなければならないという問題がある．この問題に対応するために，我々はデータクレンジング手法を用いることがある．データクレンジング手法とは，ノイズになり得るデータを削除，または修正することでデータセット全体の品質を向上させる手法である．データクレンジング手法を用いることで，学習モデルの汎化性能の向上につながることが報告されている．

しかし，データクレンジング手法を用いてノイズになり得るデータを削除する場合，作成したデータセットからラベル付けされた学習用データを削除することになり，データ数が減少してしまうという損失が生じる．そして，データクレンジング手法を用いる場合，他の教師ラベルの候補が存在するにも関わらず，他のラベルの候補が正しいかどうか検証できないという問題点もある．他の教師ラベルの候補が正しいかどうか検証できない理由は，多数決によってラベルが付与されたデータを使用するため，他のラベルの候補があったことが分からなくなっているからである．例えば，5 票の投票結果 $(1, 0, 1, 0, 1)$ があったとする．この時，多数決の結果から教師ラベルは 1 であると判断される．そのデータがデータクレンジングによって削除されるとき，判断の基準となるのは教師ラベルに選ばれた 1 のみであり，1 以外の票，つまり 0 は判断の基準になることがない．また，データクレンジング手法を用いてノイズになり得るデータを修正する場合，人手で修正するとコストがかかってしまうという問題点もある．

以上の問題点から，我々は以下の二つの問題点を解決することを目的とした．一つ目の問題点は，データクレンジング手法を用いるとデータ数が減少してしまうことである．二つ目の問題点は，多数決を行う前の投票結果の中に有用な投票結果があったとしても無視されてしまうことである．

以上の二つの問題点を解決する方法として，我々は 1) ノイズとなり得るデータを削除しつつ，多数決後のデータ数を減らさないこと，2) 投票結果のうち多数派の意見がノイズになり得るデータだった場合，少数意見である他の候補が正しいかどうか検証することが可能なことの二つの条件を満たした方法を考えた．

そこで，我々はデータセット作成過程にデータクレンジング手法を挿入することを提案する．通常のデータクレンジングと提案手法の違いは，データセット作成後にデータクレンジングを行うか，データセット作成過程でデータクレンジングを行うかである．通常のデータクレンジングは，以下の手順でデータクレンジングを行う．

（1） 多数決によって教師ラベルを決定し，教師ラベルが付与されたデータセットを用意する．

（2） データセットに対してデータクレンジングを行い，ノイズになり得るデータを削除する．

それに対して，提案手法は，以下の手順でデータクレンジングを行う．

（1） 投票結果のデータに対してデータクレンジングを行い，ノイズになり得るデータを削除する．

（2） 削除されなかったデータを用いて多数決を行い，教師ラベルを決定する．

以上で示したように提案手法はデータクレンジングを行うタイミングが異なる．我々はデータクレンジングを行うタイミングを変更することで，問題点を解決するための二つの条件を満たした手法になると考えた．

我々が 1) の条件を満たしていると考えた理由は，投票データに対してデータクレンジングを行うため，あるデータに対して集めた投票結果のうちどれか一つでも削除されなかった場合は教師ラベルを決定できるためデータ数が減少しないからである．我々が 2) の条件を満たしていると考えた理由は，多数派意見がノイズになり得るデータと判断された場合，多数派の投票データはすべて削除され，残った少数派の投票データで多数決を取ることができるからである．

本論文では，何もしない場合，通常のデータクレンジングを行った場合，提案手法を用いた場合で比較実験を行った結果を報告する．実験の目的は提案手法を用いたデータクレンジングを行うことでデータセットの品質が向上し，学習モデルの性能向上に貢献することを確認することである．データクレンジングを行う際のスコアリングに外れ値検出手法 [1] を使用し，削除率を 0%から 50%までを 5%刻みで変化させて実験を行った．その結果，通常のデータクレンジングと比較して提案手法は多数決後のデータ数が減らず，通常のデータクレンジングと同様に Accuracy が向上することを確認した．

本論文の貢献は，データセット作成時にデータクレンジングを行うことで，データ数の減少を防ぐことが可能であり，少数意見が正しいかどうか検証することができるため，ラベルを再度付与するコストがかからない手法を提案したことである．

2 関 連 研 究

データセット内の教師データにノイズが存在する場合，汎化性能が低下してしまう．この問題を解決する方法は大きく分けて二つある．

一つ目はデータクレンジングを行い，ノイズである可能性の高いデータを削除，またはラベルを修正したものを学習に使用する手法である．データクレンジングを行う方法として，提案手法と同様に外れ値検出を用いる手法以外にも，分類器の予測結果を利用する手法 [2] や，k-NN を用いた手法 [3]，Markov Random Field（MRF）を用いた手法 [4]，Area Under The Margin (AUM) Statistic を用いた手法 [5]，データ品質監視付き変分オートエンコーダ (AQUAVS) を用いた手法 [6] などがある．また，データクレンジングを逐次的な処理に変更し，ラベルノイズが疑われるデータを学習過程の中で除去する，または修正する手法 [7] [8] も提案されている．

二つ目は，データセットにラベルノイズがあることを前提とした学習モデルを設計することで，ラベルノイズに左右されない汎化性能の高いモデルを構築する手法である．半教師ありSVM を用いた手法 [9] や，ニューラルネットワークを用いた手法 [10] などが提案されている．また，ラベルノイズに強い，ロバスト性の高いモデルを構築することができる手法として，損失関数を補正する手法 [11]，学習モデルを汎化させるために重要なパラメータと重要でないパラメータに分け更新ルールを変更する手法 [12] などが提案されている．包括的なサーベイは文献 [13] を参照されたい．

我々が本論文で提案する手法は上記の手法のうち，一つ目のデータクレンジングの手法である．本論文の貢献は，データセット作成時にデータクレンジングを行うことで，データ数の減少を防ぐことが可能であり，少数意見が正しいかどうか検証することができるため，ラベルを再度付与するコストがかからない手法を実現したことである．データクレンジングを逐次的な処理に置き換える手法を除いたその他のデータクレンジング手法は，すでにラベル付けが終了しているデータセットに利用されるものである．それに対して提案手法ではクラウドソーシングにてラベルの投票を行った際の投票データに対してデータクレンジングを行う．この違いによって，データクレンジング後にラベルを付与したいデータの投票データが 1 つでも残っていれば，ラベルを付与したいデータは削除されないず，ラベルを付与することができる手法を実現した．

3 提 案 手 法

クラウドソーシングを用いたデータセットの作成は以下のように行われる．まず，ラベル付けを行う N 個のデータを用意する．このとき，N 個のデータのうち x 個目のデータを d_x と表記する．クラウドソーシングを用いて n 人の作業者に d_x のラベルを予想し，投票してもらう．このとき，d_x に対して a 人目の作業者の投票結果を $l(d_x, a)$ と表記する．その投票結果 $l(d_x, 1)$ から $l(d_x, n)$ を用いて多数決を行い，d_x に付与するラ

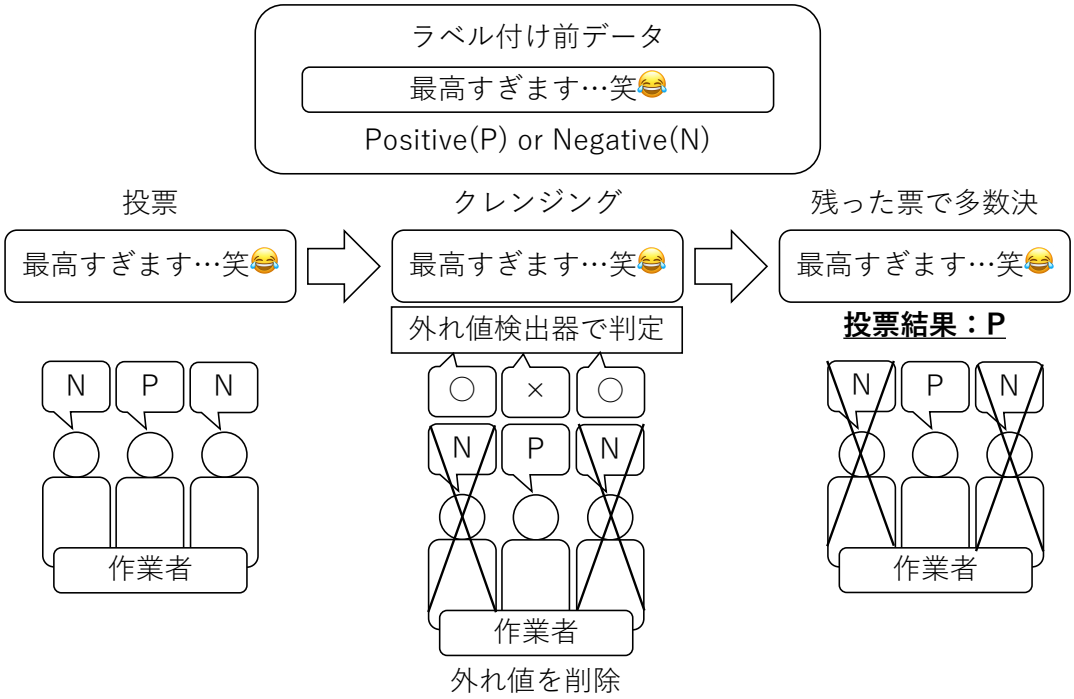

図 1　提案手法の概略図

ベル l'_x を決定する．このとき，多数決を行い作成された d_x と l'_x の組のデータセットを D_m とする．以上の手順でラベル付けが行われ，ラベル付きのデータセットが作成される．

通常のデータクレンジングは以下の手順でデータクレンジング，学習モデルの構築を行う．ラベル付けが終了した後にデータクレンジング手法を用いて l'_x がノイズになり得るかどうか判断し，ノイズになり得るデータを削除する．ノイズになり得るデータを削除したデータセットで学習モデルを構築する．以上の手順を用いることで，学習時にノイズが減り，学習モデルの性能が向上するとされている．

1 章でも述べたように，この手法には二つの問題点があると考える．一つ目は，d_x と l'_x の組が妥当でないと判断された場合，d_x は削除され，データ数が減少してしまうことである．二つ目は，多数決後のデータを参照する手法であるため，多数決を行う前の投票結果の中に有用な投票結果があったとしても無視されることである．

この 2 点の問題点を解決するために提案手法では以下の手順でデータクレンジング，学習モデルの構築を行う．多数決を行う前にデータクレンジング手法を用いて $l(d_x, a)$ が妥当かどうかを判断し，削除する．その後に削除されなかった投票結果を用いて多数決を行い，d_x に付与するラベル l_x を決定する．このとき，多数決によって作成された d_x と l_x の組のデータセットを D'_m とする．そして，D'_m で学習モデルを構築する．以上の手順を用いることで，$l(d_x, a)$ のうち一つでも削除されない投票結果が存在すれば，d_x に付与するラベル l_x を決定することが可能である．よって，提案手法は，一つ目の問題点であるデータ数が減少する点，二つ目の問題点である投票結果一つ一つを考慮できない点を解決した手法であると考える．

提案手法は以下の手順でデータクレンジングを行う．提案手法の大まかな流れを図 1 に示す．まず，クラウドソーシングを行い，N 個の d_x と $l(d_x, 1)$ から $l(d_x, n)$ の情報をもつデータからなるデータセット D_b を用意する．次に，外れ値検出器を用いて D_b の d_x と $l(d_x, a)$ の組がノイズになり得るかどうかの判定を行う．外れ値検出器については 3.1 節で述べる．そして，外れ値検出器によりノイズになり得ると判断された d_x と $l(d_x, a)$ の組を削除する．このとき，ノイズになり得ると判断された d_x と $l(d_x, a)$ の組を D'_m から削除した後のデータセットを D_c とする．最後に，削除されなかった残りの投票結果を用いて多数決を行い，d_x に付与するラベル l_x を決定する．このとき，多数決によって作成された d_x と l_x の組のデータセットが D'_m である．以上の手順でデータクレンジングを行い，データクレンジング後のデータセット D'_m を学習に使用する．

3.1　外れ値検出手法

提案手法を適用するにあたり，データクレンジングの手法に意図的な過学習を利用した外れ値検出 [1] を利用した．今回使用した外れ値検出手法はデータセットを学習させたニューラルネットワークを用いて，データセットのうち，データ一つをさらに学習させた際の重みの変化を使用する手法である．以下の節で詳しく説明する．

3.1.1　事 前 学 習

ニューラルネットワークを用いて D_m を学習させた基準モデルを構築する．この時点ではデータクレンジングは行っていない為，データセットには D_m を使用することに注意されたい．D_m のデータを訓練用，検証用，テスト用のデータセット（D_{train}，D_{val}，D_{test}）に分割して学習を行う．

3.1.2 個別の学習

まず，D_b から d_x と $l(d_x, a)$ の組一つを取り出す．次に，d_x を入力データ，$l(d_x, a)$ を教師データとして，基準モデルをファインチューニングする．この時，ニューラルネットワークの重みは最終層のみ更新する．ファインチューニングにより，d_x を $l(d_x, a)$ であると分類できた時点の最終層の重みを d_x と $l(d_x, a)$ の組の特徴量 $f(d_x, a)$ とする．以上の処理を D_b のデータがなくなるまで繰り返す．

3.1.3 スコアリング

D_b から取り出した d_x と $l(d_x, a)$ の組に対応する特徴量 $f(d_x, a)$ を使用し，マハラノビス距離によるスコアリングを行う．特徴量 $f(d_x, a)$ のマハラノビス距離を $M(f(d_x, a))$ とする．求めた全ての $M(f(d_x, a))$ のうち，最大の距離を M_{max} とする．$M(f(d))/M_{max}$ を計算することで求められる値を d_x と $l(d_x, a)$ の組のスコアとする．上記のスコアは 0 から 1 の値をとり，1 に近いほど外れ値である可能性が高いスコアである．

4 評価実験

提案手法が有効であることを示すため，評価実験を行った．データクレンジングを行わない場合，通常通り投票データを集計したデータセットに対してデータクレンジングを行う場合，提案手法を用いる場合の 3 パターンで学習を行い，Accuracy の比較を行った結果を報告する．実験に使用するデータセットには「笑」が付いた文章の感情分析を行うために実施したクラウドソーシングの結果を用いた．また，学習には BERT [14] を用い，損失関数に CrossEntropyLoss，最適化手法に Adam を用いた．

4.1 使用データ

実験では「笑」が付いた文章の感情分析を行うために実施したクラウドソーシングの投票結果を用いた．投票データを基にラベルを付けたいデータにラベルを付与する．ラベルの候補はネガティブ，ニュートラル，その他，攻撃的，非攻撃的，判定不可，ポジティブ，自虐，ポジティブ＋ネガティブの 9 種類である．これらのラベルのうち，ネガティブを細分化したものが攻撃的，非攻撃的，判定不可，自虐の 4 種類である．その他のラベルは想定外のデータのために用意されている．そこで，今回の実験では，ネガティブ，その他，判定不可に投票されたデータは削除したうえで実験を行った．「笑」が付いた文章の感情分析を行うために実施したクラウドソーシングの投票結果の場合，5 人の投票を基に多数決を取り，付与するラベルを決定する．「笑」が付いた文章の感情分析を行うために実施したクラウドソーシングの投票結果には投票データが合計で 250,354 件ある．一つのデータあたり 5 票の投票が行われているため，集計すると約 50,000 件のラベル付けされたデータが作成されることになる．しかし，ラベル付けを行いたいデータ一つに対して投票データが 5 票集まっておらず，集計できないデータがある．そのため，集計可能な 5 票集まっている投票データのみを使用した．その結果，使用可能な投票データは 137,505 件になった．票数が足りない原因は，ネガティブ，その他，判定不可を削除したこと，もともと票数が足りていないことがあげられる．よって，評価実験は集計が可能な 137,505 件のデータを対象に行う．

4.2 実験手順

提案手法が有効であることを示すため，データクレンジング無しで学習を行った場合，通常のデータクレンジングありで学習を行った場合，提案手法のデータクレンジングありで学習を行った場合の 3 パターンの実験を行う．まず，多数決のルールを説明する．多数決は最初にポジティブ，ネガティブ，ニュートラル，ポジティブ＋ネガティブの 4 種類で多数決を行う．この時複数のラベルの票数が同じになった場合そのデータは削除することとする．上記の 4 種類で多数決を行った結果，ネガティブと判定されたものについては追加で攻劇的，非攻撃的，自虐の 3 種類の票で多数決を行う．この時複数のラベルの票数が同じになった場合，そのデータは削除することとする．

データクレンジング無しで学習を行う場合は，以下の手順で実験を行う．

手順 1　投票結果をもとに多数決によって教師ラベルを決定しデータセットを作成する．その結果データ数は 23,225 件になった．

手順 2　作成したデータセットを 訓練データ：検証データ：テストデータ = 8: 1: 1 に分割し，学習を行う．

手順 3　学習終了後テストデータを用いたときの Accuracy を性能の評価に利用する．

データクレンジングありで学習を行う場合は，以下の手順で実験を行う．

手順 1　データクレンジング無しの場合と同様に投票結果をもとに多数決をとり，23,225 件のデータからなるデータセットを作成する．

手順 2　作成した 23,225 件のデータを外れ値検出器に入力し，ノイズになり得るデータを削除するためのスコアをつける．

手順 3　付けられたスコアをもとにノイズである可能性が高い上位 5%から 50%（5%刻みで変更して実験）を削除する．

手順 4　削除後のデータセットを使用し，訓練データ：検証データ：テストデータ = 8: 1: 1 に分割し，学習を行う．

手順 5　学習終了後テストデータを用いたときの Accuracy を性能の評価に利用する．

提案手法で学習を行う場合は，以下の手順で実験を行う．

手順 1　投票結果 137,505 件を外れ値検出器に入力し，ノイズになり得るデータを削除するためのスコアをつける．

手順 2　付けられたスコアをもとにノイズである可能性が高い上位 5%から 50%（5%刻みで変更して実験）を削除する．

手順 3　削除後の投票データを使用して多数決を取る．このときの多数決は 4.1 節で述べた多数決とは異なり，投票数が 5 個揃っていない場合でも残っている投票データのみ

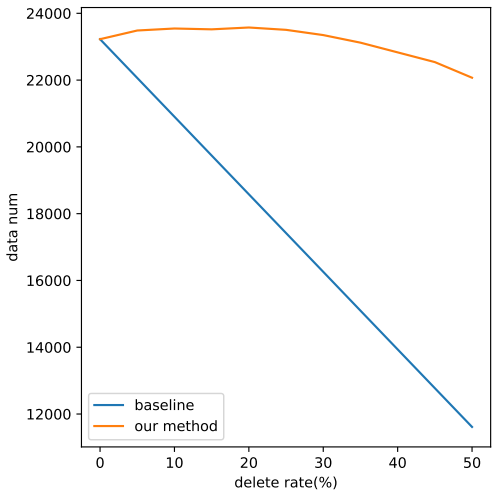

図 2　データクレンジングの削除率とデータ数の変化

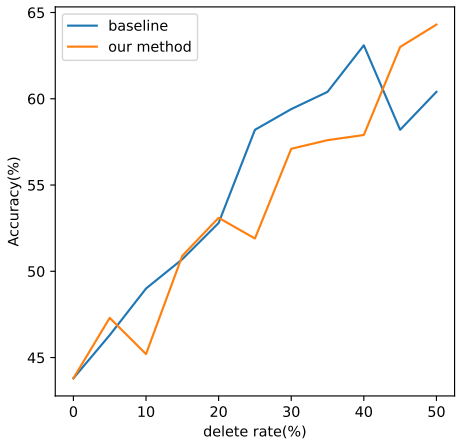

図 3　データクレンジングの削除率と Accuracy の変化

で多数決を取る．

手順 4　多数決後のデータセットを使用し，訓練データ：検証データ：テストデータ = 8: 1: 1 に分割し，学習を行う．

手順 5　学習終了後テストデータを用いたときの Accuracy を性能の評価に利用する．

データクレンジングありで学習を行う場合，提案手法で学習を行う場合のデータ数の変化については 4.3 節で述べる．これら 3 パターンの比較実験により，提案手法はデータクレンジング後のデータを減らすことなくデータクレンジングと同様の効果を得ることができる手法であることを示す．そして，提案手法によってどの程度データ数を減らすことなくデータクレンジングが可能かどうかを調べるため，削除する割合とデータクレンジング後のデータ数の関係を調査する．最後に，実際にどのようなデータの教師ラベルが変更されたかを確認し，その教師ラベルの変更が妥当なものだったかどうかを議論する．

4.3　実 験 結 果

図 2，図 3 に結果を示す．図 2 に削除率を増加させた際のデータ数の変化を示す．横軸が削除率，縦軸が多数決後のデータ数である．青い線がデータクレンジングありで実験したときの結果（baseline），オレンジの線が提案手法で実験したときの結果 (our method) を示している．図 2 からデータクレンジングありで実験 (baseline) した場合は削除率を減らしていくと，作成済みのデータセットから多数決後のデータが削除されていくため，直線的に多数決後のデータ数が減少していくことが確認できる．それに対して提案手法で実験 (our method) した場合は削除率を減らしていくと徐々に多数決後のデータ数が増加し，その後緩やかに減少していくことが確認できる．多数決後のデータ数が増加する理由は投票データが削除されたことによって票数が同率だったために削除されていた投票データのラベルが決定するためである．

また，削除率が 0%，つまりデータクレンジングを行わない場合の多数決後のデータ数が 23,225 件であった．データクレンジングありで実験した場合は多数決後のデータ数が最大で半分の 11,613 件にまで減少してしまう．それに対して提案手法で実験した場合は多数決後のデータ数が最大で 23,575 件に増加し，最小で 22,071 件に減少することが分かった．

図 3 に削除率を変化させた際の Accuracy の変化を示す．横軸が削除率，縦軸が Accuracy である．青い線がデータクレンジングありで実験したときの結果（baseline），オレンジの線が提案手法で実験したときの結果 (our method) を示している．図 3 からデータクレンジングありで実験した場合も，提案手法で実験した場合もどちらも削除率の増加と共に Accuracy が向上することが確認できた．データクレンジングありで実験した場合は最大で 19.3%，提案手法で実験した場合は最大で 20.5% の Accuracy の増加が確認できた．

4.4　考　　察

4.3 節でも述べたように，提案手法は多数決後のデータ数が減らず，Accuracy の向上についても通常のデータクレンジングと変わらないことが分かった．よって，提案手法は 1 章で述べた，1) ノイズとなり得るデータを削除しつつ，多数決後のデータ数を減らさないこと，2) 投票結果のうち多数派の意見がノイズになり得るデータだった場合，少数意見である他の候補が正しいかどうか検証することが可能なことの二つの条件を満たしたうえで通常のデータクレンジングと同様にモデルの性能向上につながる手法であると考える．

また，投票データを削除することで少数意見が通るようになることを期待したが，その副産物として票数が同率のためにラベルの付与が行えないデータにラベルを付与することが可能になることが分かった．この効果によって通常のデータクレンジングでは学習に使用可能な多数決後のデータ数が減る一方だったのに対して，提案手法では多数決後のデータ数が増加するという結果を得ることができた．そして，データの 50%を削除した場合，通常のデータクレンジングであれば多数決後のデータ数が元のデータ数の半分になるのに対し，提案手法では元のデータ数から 5%減少するのみであることが分かった．よって本研究の目的の一つである，データ数を減らさないデータクレ

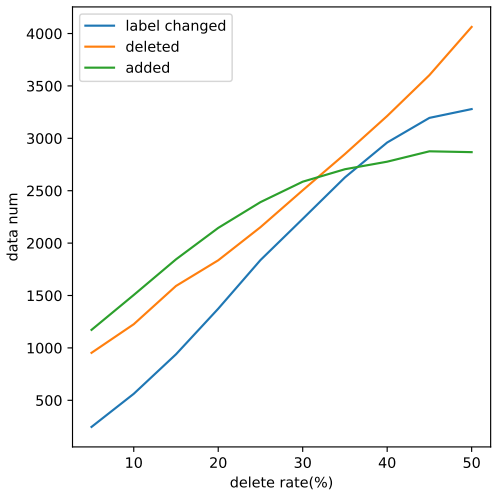

図 4　提案手法の削除率とラベルが変化したデータ数，削除されたデータ数，追加されたデータ数の変化

ンジングを実現できたと考える．また，少数意見が通るようになったために性能が向上しているかどうかは実際にどのようなラベルの変化があったのかを確認する必要があると考える．

そして，図 3 から，通常のデータクレンジングと提案手法はどちらも Accuracy の向上につながる手法であることが確認できた．提案手法の Accuracy が増加した原因は新たにラベルを付けることができた同率だったデータのおかげなのか，データクレンジングと同様の効果が得られているからなのかはまだ不明である．しかし，少なくとも通常のデータクレンジングと同様の効果が得られることは確認できた．

そこで，データクレンジング前の多数決後のデータとデータクレンジング後の多数決後のデータの違いを確認することにした．その結果と考察については次節で述べる．

4.5　データクレンジング前後の比較

まず我々は提案手法の削除率を 5%刻みで 0%から 50%に変化させたときの以下に示す三つの項目の変化を調べた．一つ目は提案手法によって多数決後の教師ラベルが変化したデータ数 (label changed)，二つ目はデータクレンジングをしない場合は存在したが，提案手法によって削除されたデータ数 (deleted)，三つ目は提案手法によって同率だった投票数に変化が起こり，教師ラベルを決定できたために追加されたデータ数 (added) である．これら三つの推移を図 4 示した．横軸が削除率，縦軸がそれぞれのデータ数を示している．

図 4 から，提案手法の削除率を上昇させることによってラベルが変化したデータ数，削除したデータ数，追加されたデータ数は全て増えていくことが分かった．今回の実験では，削除率を 30%より高く設定すると追加されたデータ数が削除したデータ数を下回る．つまり，30%以下であれば多数決後のデータ数を減らさないデータクレンジングを実現できているといえる．一方で 30%以上の削除率を設定すると，データクレンジングを行わない場合と比較して徐々に多数決後のデータ数が減少していくことが分かった．よって今回使用したデータセットでは削除率を 30%程度に設定することで，本研究の目的であるデータ数を減らすことなく，Accuracy の向上が期待できるデータクレンジングを行うことが可能であるといえる．

提案手法を適用した場合（削除率 50%）の多数決後のデータの変化を表に示す．表 1 にはラベルが変化したデータの例，表 2 には削除されたデータの例，表 3 には追加されたデータの例を示す．それぞれ良かった例と悪かった例を一つずつ示してある．それぞれについて言及する．

表 1 の良かった例について述べる．多数決後のラベルはもともとニュートラルとなっていたが，提案手法によってポジティブ＋ネガティブに変化した．文章を読んでみると「なんか様になった (笑)」の前後に両親が亡くなったという情報があり，ネガティブな要素があると推測される．そしてお骨を奉るのに普通は使わないキーボードスタンドを使用したのに様になったのが面白かったという一面もある．それらを考慮するとポジティブ＋ネガティブの方がラベルとして妥当だったと考えられ，提案手法によって上手くラベルが修正された例であると考えられる．

表 1 の悪かった例について述べる．多数決後のラベルはもともとニュートラルとなっていたが，提案手法によって非攻撃的に変化した．文章を読んでみると「イケボ配信者のイケボ (笑)」と煽っているように取れる文面である．そのため，付与されるべきラベルは攻撃的であると考えられる．今回の実験で得られた結果は非攻撃的であるためラベルが間違っている．しかし，これは提案手法が悪いのではなく，そもそも，投票結果に攻撃的という票が入っていない．そのため．元のニュートラルよりは近いネガティブなラベルになったものの，正しいラベルにはならなかったと考えられる．

表 2 の良かった例について述べる．多数決後のラベルはもともとポジティブとなっていたが，提案手法によって削除された．意味の取り方によるかもしれないが，怖いからひとりでは行けないという意味で読んでも，恥ずかしくて行けないという意味でも自虐になる．そして，このデータに対する投票結果の中に自虐というラベルは存在していない．よって，このデータは誤ったラベルが付与されたデータであり，削除されても良かったデータであると判断できる．

表 2 の悪かった例について述べる．多数決後のラベルはもともとポジティブとなっていたが，提案手法によって削除された．「第 2 子の胎動を観測」とあるので，喜んでいることが想像できるうえ，夫が寝ているから Twitter でという意図が見て取れる．よって付与されるべきラベルはポジティブであり，削除されるべきではないデータであると判断できる．

表 3 の良かった例について述べる．多数決後のラベルはもともと同率票によりラベルを付与できなかったが，提案手法によって非攻撃的であると判断された．昼ごはんが甘いものだったからコーヒーが飲みたい．しかし，コーヒーがないという状態が推測されるため，ネガティブな感情かつ，攻撃的ではないと予想される．よって付与されるべきラベルは非攻撃的である．提案手法によってニュートラルの票が削除されたことにより，非攻撃的であると判断可能になったと推測される．よって，

表 1　提案手法の適用によって多数決後のラベルが変化したデータの例

	文章の例	ラベルの変化	理想のラベル	実際の投票結果
良かった例	日常が　心の隙間　埋めてゆく母のお骨を部屋に奉るのに 丁度いい台がなかったので楽器のキーボードスタンドに 白い布団カバーを敷いて代用。なんか様になった (笑) 仕事が容赦ないのは逆にありがたい。 忙しさが心の隙間を埋めてくれる。 親が死んでも働く兄妹であります。#09 月 27 日 #3 行日記	ニュートラル ↓ ポジ＋ネガ	ポジ＋ネガ	ポジ＋ネガ：2 ニュートラル：3
悪かった例	YouTube の広告とかで見る勘違いイケボ配信者のイケボ (笑) すごく苦手なんですよ	ニュートラル ↓ 非攻撃的	攻撃的	ニュートラル：2 非攻撃的：1 ポジ＋ネガ：1 ポジティブ：1

表 2　提案手法の適用によって多数決後のデータから削除されたデータの例

	文章の例	ラベルの変化	理想のラベル	実際の投票結果
良かった例	実は、息子が出来たら恐竜博物館に一緒に行くのが夢です w 一人じゃ行けないから笑	ポジティブ ↓ 削除	自虐	ポジティブ：3 ポジ＋ネガ：1 ニュートラル：1
悪かった例	ハッ… !!!!! ただ今第 2 子の初胎動を観測しました… 夫寝てるからツイートしとこ笑	ポジティブ ↓ 削除	ポジティブ	ポジティブ：5

表 3　提案手法の適用によって多数決後のデータに追加されたデータの例

	文章の例	ラベルの変化	理想のラベル	実際の投票結果
良かった例	お昼が甘甘過ぎてブラックコーヒーが飲みたい…無い（笑）	削除 ↓ 非攻撃的	非攻撃的	非攻撃的：2 ニュートラル：2 ポジティブ＋ネガティブ：1
悪かった例	まぁ、モテるに決まってるよねぇ。だってジャニーズだもん (笑)	削除 ↓ 攻撃的	ポジティブ	ポジティブ：2 ニュートラル：2 攻撃的：1

提案手法によって上手くラベル付けができた例であると考えられる．

表 3 の悪かった例について述べる．多数決後のラベルはもともと同率票によりラベルを付与できなかったが，提案手法によって攻撃的であると判断された．「まぁ、モテるに決まってるよねぇ。だってジャニーズだもん (笑) 」という文面をそのまま見るとほめているためポジティブであると判断されるべきである．しかし，提案手法によってポジティブ，ニュートラルの票が削除されてしまい，攻撃的であると判断されてしまっている．よって提案手法によるラベル付けがうまくいかなかった例であると考える．

以上に示したように多数決後のラベルが提案手法によって変化したデータは，正しく削除されたもの，正しいラベルに変化したものもあれば，正しく削除されたもの，正しいラベルに変化したとは言えないものもあることが分かった．文章を文面のまま受け取るか，裏の部分を考えるかでも回答が変わるため，意味の取り方によっては正しいとも，間違っているとも取れるような投票データも確認された．

また，投票データに依存する問題点があることも分かった．例えば，表 2 の悪かった例のようにそもそも投票データ内に付けてほしいラベルが存在していない場合，このデータのラベルは間違うことしかできない．この問題点はクラウドソーシングによる投票でデータセットを作成する際の問題であるため，提案手法特有の問題ではないが，解決する必要があると考える．

5　おわりに

本論文で我々はデータクレンジングをデータセット作成過程に挿入することで，データ数の減少を抑えたデータクレンジングを提案した．データクレンジングをデータセット作成過程に挿入することで作業者の作業結果を全て確認することができ，データクレンジングによって他のラベルの候補が無視されてしまう問題の解決した．

そして，提案手法でもデータクレンジングと同様の効果が得られることを示すため，データクレンジングを行わない場合，通常のデータクレンジングによってノイズになり得るデータの削除を行った場合，提案手法を適用した場合の比較を行った．その結果，提案手法は通常のデータクレンジングと同様に Accuracy の向上を図ることができる手法であることを確認できた．また，通常のデータクレンジングでは多数決後のデータ数が 50%減少する場合でも，提案手法を用いることで，多数決後のデータ数は 5%しか減少しないことを確認した．

以上の結果から，提案手法は 1 章で述べた条件，1) ノイズとなり得るデータを削除しつつ，多数決後のデータ数を減らさ

ないこと，2) 投票結果のうち多数派の意見がノイズになり得るデータだった場合，少数意見である他の候補が正しいかどうか検証することが可能なことの二つの条件を満たすことができた．

以上のことから我々はデータセットを公開する際には教師ラベルを付与したデータを公開するのではなく，付与する前のデータを公開すべきであると主張する．本論文で使用したデータセットを例とするのであれば，（テキスト，教師ラベル）の組を提供するのではなく，（テキスト，投票結果 1，…，投票結果 5）を公開すべきだと考える．理由は提案手法を適用してそれぞれ任意の削除率でクレンジングを行い，学習を行うことができるためである．投票データ自体を公開することで提案手法が適用可能となり，データ数の減少を防ぎつつ，データクレンジングと同様の Accuracy の向上を期待できると考える．

今後の展望として，データクレンジング手法で削除することのできない誤ったラベル付きデータに対してどのような対策が考えられるか検討していきたいと考えている．まず最初に検討するべきなのは，今回使用していないデータクレンジング手法を組み合わせることである．また，削除できない原因として，純粋に間違っているものを検出できない場合や，データセット全体に存在するバイアスによって検出できない場合など，様々な原因が考えられる．データセット全体にバイアスが存在している場合，既存の手法はデータセット全体の分布を基にするものがほとんどであるため，検出できないと考えられる．そのため，新たな手法を考える必要があると考える．

謝辞　本研究の一部は JSPS 科研費 19H04218 および越山科学技術振興財団の助成を受けたものです．

文　献

[1] 三島惇也, 鈴木優. 意図的な過学習によるパラメータの変化を用いた外れ値検出. 第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM2022), pp. H34–1, 2022.

[2] Jaree Thongkam, et al. Support vector machine for outlier detection in breast cancer survivability prediction. In *Advanced Web and Network Technologies, and Applications*, pp. 99–109, Berlin, Heidelberg, 2008.

[3] Sarah Jane Delany and Pádraig Cunningham. An analysis of case-base editing in a spam filtering system. In *Advances in Case-Based Reasoning*, pp. 128–141, 2004.

[4] Karishma Sharma, Pinar Donmez, Enming Luo, Yan Liu, and I. Zeki Yalniz. Noiserank: Unsupervised label noise reduction with dependence models. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pp. 737–753, Cham, 2020.

[5] Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 17044–17056. Curran Associates, Inc., 2020.

[6] Vaibhav Pulastya, Gaurav Nuti, Yash Kumar Atri, and Tanmoy Chakraborty. Assessing the quality of the datasets by identifying mislabeled samples. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, p. 18–22, New York, NY, USA, 2022.

[7] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi-Phuong-Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. SELF: learning to filter noisy labels with self-ensembling. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[8] Stefano Teso, Andrea Bontempelli, Fausto Giunchiglia, and Andrea Passerini. Interactive label cleaning with example-based explanations. In *Advances in Neural Information Processing Systems*, Vol. 34, pp. 12966–12977, 2021.

[9] Lorenzo Bruzzone and Claudio Persello. A novel context-sensitive semisupervised svm classifier robust to mislabeled training samples. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 47, No. 7, pp. 2142–2154, 2009.

[10] Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew Nokleby. An effective label noise model for DNN text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3246–3256, Minneapolis, Minnesota, June 2019.

[11] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin Mcguinness. Unsupervised label noise modeling and loss correction. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 312–321. PMLR, 09–15 Jun 2019.

[12] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International Conference on Learning Representations*, 2021.

[13] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–19, 2022.

[14] Jacob Devlin, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc.NAACL-HLT, Volume 1*, pp. 4171–4186, 2019.