CF-002

ノイズに悩まされないで：シャドウデータに基づくラベルノイズ検出方法
Don't Let Noise Bother You: A Shadow-Instance-Based Framework for Label Noise Detection

鄭　彎彎†
Wanwan Zheng

1. はじめに

機械学習は，データからパータンや規則性などを学習したモデルで未知のことを予測する．学習用のデータそのものの誤りがあれば，モデルの質に大きな影響を与える．これらの誤りはノイズと呼ばれる．ノイズはラベルノイズと変数ノイズに分けられるが，前者は学習のターゲットとしてより有害であると指摘されている．具体的には，データの規則性が捉えにくくなることによって，モデルの複雑性が増大し，全体的な精度が低下する．その結果，誤った情報が解釈に反映されることになる．しかし，現実世界では，制御された環境でも，獲得データには少なくとも 5%程度の誤りが含まれることが想定され，ラベルノイズの存在は回避できない．この問題を軽減するため，データ中心のノイズ検出（label noise detection）とアルゴリズム中心のロバストな学習アルゴリズム構築（label noise-tolerant methods）という二つの側面から研究が行われている．本研究は，ラベルノイズの検出に注目する．

- **ラベルノイズ検出と個体選択**

ビッグデータ時代における電子化の促進により，あらゆる情報が瞬時に取得できるようになった．そこで，高精度の学習モデルを作成するためにはデータの量からデータの質に注目するようになっている．データの質を落とさずにデータの量を減らす方法は，ラベルノイズ検出（label noise detection）と個体選択（instance selection）があるが，この二つは本質的に目的が異なる．

ラベルノイズ検出は，所属クラスに対して異質な属性をもっているサンプルを検出する．例えば，クラス A のラベルが付与されているが，実際にはクラス B の属性を顕著に表しているような **X** がそれに該当する．個体選択は，ノイズと情報重複のサンプルを検出し，最後に代表的な個体しか残さない．例えば，正常なサンプルであり，高い相関をもっている **X** と **Y** のうち，どちらかを削除しても境界線に影響しないと判断された場合に **X** と **Y** の中から一つを取り除くことを言う．一般的に個体選択はラベルノイズ検出と比べ，検出率が高い．後者では学習モデルの精度に重点が置かれているが，前者では学習スピードを加速することに重点が置かれているため，精度を若干犠牲にするケースも度々ある．

- **ラベルノイズ検出の手順**

ラベルノイズ検出を含む一般的な学習手順を図 1 に示す．既存のラベルノイズ検出方法では，ラベルノイズ検出段階でノイズであるかどうかが明確に判別できないため，多数決投票または閾値設定でノイズの除外/訂正を行う．

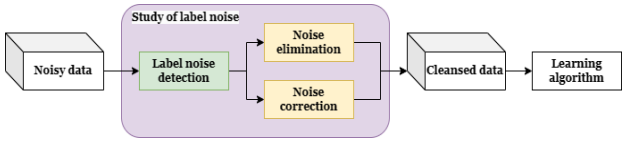


図 1　ラベルノイズ検出を含む一般的な学習手順

†名古屋大学大学院人文学研究科　Graduate School of Humanities, Nagoya University

Gupta and Gupta（2019）は，1993 年から 2018 年までの間に発表されたラベルノイズ検出に関する研究をレビューした．その結果，過去 20 年間に数多くのアンサンブルベースの方法によるラベルノイズの検出精度が最も高かったことがわかった．有効性と汎用性を備えたアンサンブルベースのラベルノイズ検出方法は現時点でも主流になっている．その後，Nematzadeh ら（2020）はアンサンブルフィルターと距離ベースフィルターを含むラベルノイズ検出方法を提案し，Feng ら（2020）は四つのノイズ検出指標，三つのアンサンブル分類器を用いて判別精度を最大にするアダプティブ方法を提案した．Yao ら（2022）は， XGBoost やランダムフォレストなど五つをベース分類器とし，ロジスティック回帰をブレンド方法にしたスタッキングにより，誤判別されたサンプルの所属確率をノイズスコアとして算出する方法を提案した．

これらのようなアンサンブル構造の方法は，高い検出率を確保できるが，使用したベース分類器に影響されること，計算コストが高いこと，そして閾値の設定の仕方が課題になっている．

そこで，本研究は，分類器を使わないアンサンブル構造のラベルノイズ検出方法 ShadowN（a Shadow-instance-based framework for label Noise detection）を提案する．ShadowN の特徴として，速い計算スピードと閾値なしという二つが挙げられる．

2. ベースライン：Confident learning

機械学習では，ベンチマークデータを用いてアルゴリズムの性能を評価するのが一般的である．これは，ベンチマークデータはクリーンであるという仮定の上で行われている．しかし，Northcutt ら（2021）は「最もよく使用されている機械学習のベンチマークデータにおける数百万のラベル誤りが存在している」を報告し，ラベルノイズの検出，訂正，学習機能を有する confident learning（CL）を提案した．CL は国際会議 ICML2020 で初めて発表され，その応用は NeurIPS2021 の best paper として選ばれた．さらに，CL の実装 Cleanlab[1]は Google，Amazon，ORACLE などに使用され，新しい機械学習モデルを提案するよりも優先度が高い手法であると言われている．

CL は，データと任意のモデルの出力確率を用いて，真のラベルと与えられたラベルの同時分布を推定することでどのサンプルのラベルが間違っているのか，ラベルノイズはどのような特徴をもっているのか，どのクラスはどのラベルに間違えられやすいのかなどの問題を解く．ここで，モデルの出力確率が最も大きいものを真のラベルに採用する．報告によると，CIFAR-10 において，ノイズ割合が 0.2 と 0.4 のとき，検出適合率はそれぞれ 0.7，0.8 程度になっ

[1] https://cleanlab.ai

ている．さらに，データクリーニングのあとの学習モデルの正解率は，既存方法と同等，またはそれ以上の良い結果が得られる．

CL はどのようなデータにも適応可能であるとされているが，しかし，既存の研究では画像，テキスト，音声の膨大なデータしか実装されておらず，また人工ノイズを使用している点で問題がある．

本研究では，より小規模のデータに対して提案手法と CL を比較する．人工ノイズは真のノイズと異なる性質を持つという視点から，ノイズの注入には，元データのラベルを入れ替える方法を採用する．また，前章で紹介したいくつの最新のアンサンブル構造の方法に関しては，ソースコードが公開されていないため，今回比較対象から外す．

3. シャドウデータに基づくラベルノイズ検出方法

ShadowN の構造と理論的な背景を図 2 に示す．各クラスに対して，以下の処理を行う．

ステップ 1．ランダムに特徴量を生成してシャドウデータを作成する．

ステップ 2．変数の一部をランダムサンプリング（復元抽出）し，新たなサブセットを作成する．

ステップ 3．サブセットを対象にし，各サンプルのノイズスコアを算出する．ノイズスコアは，ノイズである確率を示すため，値が大きいほど，ノイズである確率が高い．

ステップ 4．オリジナルデータのスコアとシャドウデータのスコアを比較する．

ステップ 5．ステップ 1~4 を 100 回繰り返し，データごとにシャドウデータのスコアと student's t 検定を行う．平均値が大きくかつ有意な差があるものはノイズとみなす（$\alpha = 0.01$）．

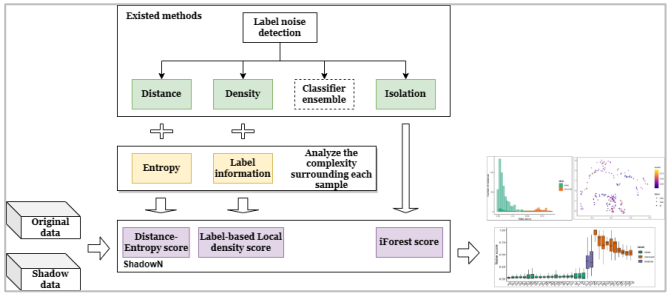


図 2　ShadowN の構造と理論的な背景

3.1 ノイズスコアの計算

この節では，ShadowN のコアの部分であるノイズスコアの算出を説明する．既存方法の距離，密度と Isolation という方法論に基づき，各サンプル周辺の複雑度を測るエントロピーとラベル情報を加え，距離–エントロピースコア，ラベル付き局所密度スコアと iForest スコアを算出する．最終的に，サンプル $\mathbf{X}$ のノイズスコアは以下のように定義する．

$$s(\mathbf{X}, k) = w_{de} \times de(\mathbf{X}, k) + w_{ld} \times ld(\mathbf{X}, k) + w_{if} \times iforest(\mathbf{X}, k)$$

$$w_{de} + w_{ld} + w_{if} = 1$$

de $(\mathbf{X}, k)$: 距離–エントロピースコア

ld $(\mathbf{X}, k)$: ラベル付き局所密度スコア

iforest $(\mathbf{X}, k)$: iForest スコア

k: 最近傍 k

- **距離－エントロピースコア**

サンプル $\mathbf{X}$ の距離－エントロピースコア de $(\mathbf{X}, k)$は，最近傍 k 個のサンプルの中に $\mathbf{X}$ と違うラベルを持つサンプルの割合 $p(\mathbf{X}, k)$に基づいて計算する．$p(\mathbf{X}, k) = 0$ の場合，de $(\mathbf{X}, k)$は 0 になる．$p(\mathbf{X}, k) = 1$ の場合，de $(\mathbf{X}, k)$は 1 になる．$0 < p(\mathbf{X}, k) < 1$ の場合，de $(\mathbf{X}, k)$は $p(\mathbf{X}, k) \times$ (1-entropy$(\mathbf{X}, k)$)になる．Entropy$(\mathbf{X}, k)$は $-\sum_{k=1}^{K} p(\mathbf{X} = k) \log p(\mathrm{X} = k)$と同じ値を取る．

- **ラベル付き局所密度スコア**

局所密度は，密度において $\mathbf{X}$ が周辺のサンプルに対しての局所的な偏差を示す．異常値検出によく使用される指標である．異常値の局所密度は実質的に低いという考え方を踏まえ，ラベルの情報を加えて所属クラスと異なるクラスにおける各サンプルの局所密度を計算する．それぞれ hit $(\mathbf{X}, h)$と dif $(\mathbf{X}, f)$で示し，h と f はサンプル数である．

$p(\mathbf{X}, k) = 0$ の場合，ld $(\mathbf{X}, k)$は 0 になる．$p(\mathbf{X}, k) = 1$ の場合，ld $(\mathbf{X}, k)$は 1 になる．$0 < p(\mathbf{X}, k) < 1$ の場合，ld $(\mathbf{X}, k)$は 1-sigmoid(hit $(\mathbf{X}, h)$/dif $(\mathbf{X}, f)$)になる．

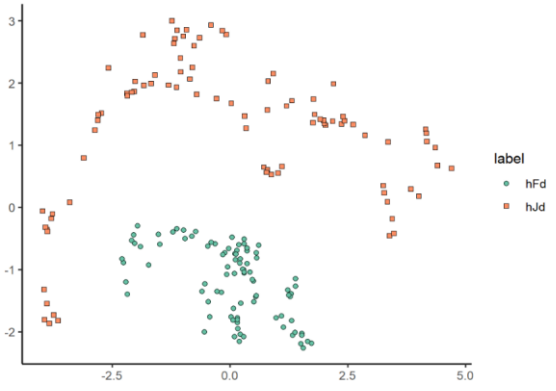
- **iForest スコア**

iForest は，Liu ら（2008）に提案された異常値検出方法であり，最も広く使用されている方法の一つである．iForest は，距離や密度の尺度でなく，孤立木（isolation tree）の概念を定義したものである．孤立木では，データに対してランダムに選ばれた変数の分割点で軸平行切断を行い，1 つのサンプルしか含まないシングルトン・ノードが生成されるまで再帰的分割を行う．異常値を含むノードは，データポイントが疎な領域に位置するため，木の深いところに至らない．そこで，葉と根の距離が異常スコアとして使用される．iForest は，複数の孤立木のアンサンブル結合であり，異なる木におけるデータポイントの経路長を平均化する．

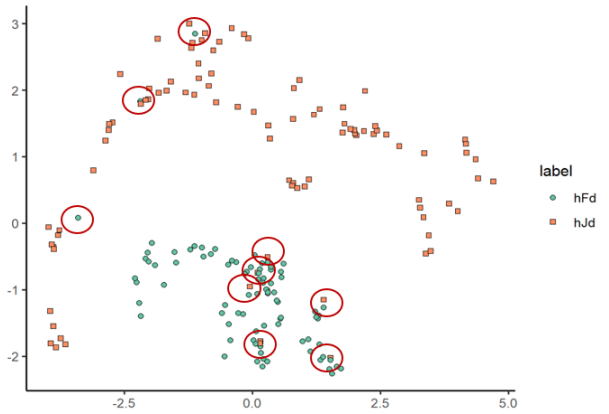
4. ShadowN の有効性

この節では，ShadowN がノイズを正しく検出できるかどうかを検証する．ベンチマークデータとして，Japanese vowels を用いた．Japanese vowels は，日本人男性話者 9 人から採取した 9 個の LPC セプストラム係数の音声データである．先述のように，ベンチマークデータの中にラベルノイズが存在する可能性があるので，データが重なっておらず明確な境界づけられる 2 つのクラス hFd と hJd のデータを用いた．サンプル数は 180 個，そのうち hFd と hJd のサンプル数は同じである；変数数は 9 個で，すべて連続データである．また，10 個のサンプルをランダムに抽出し，ラベルを入れ替え，ノイズとして使用した．元データとノイズを入れたデータについて主成分分析を行った結果を図 3 に示す．

図 4 には ShadowN が算出したサンプルごとのノイズスコアを示す．ノイズは正常サンプルより高いスコアを取得



（A）元データ



（B）ノイズを入れたデータ

図 3　主成分分析によるデータ構造

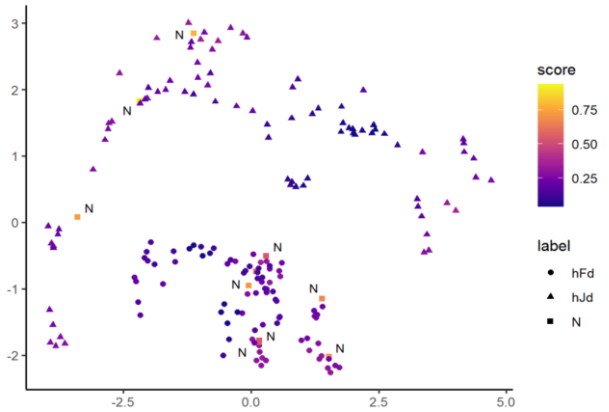


図 4　ShadowN で算出したノイズスコア

しすべて検出された．

さらに詳しく考察するため，100 回実行して得られたシャドウデータ，正常データとノイズデータのスコアの箱ひげ図を図 5 に示す．ただし，正常データの数が多いため，スコアが最も小さい 20 個のデータだけをプロットした．正常データのノイズスコアは，明らかにシャドウデータのスコアを下回った．一方，ノイズのスコアはシャドウデータのスコアを上回った．以上により，シャドウデータは正常データとノイズの判別に機能したことがわかった．

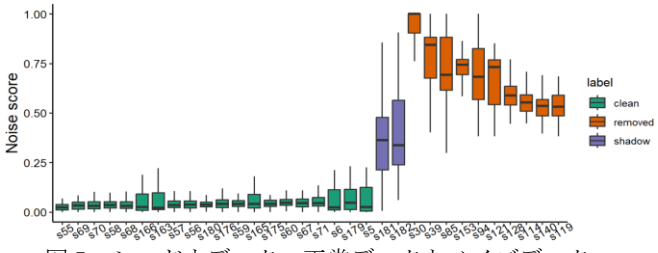


図 5　シャドウデータ，正常データとノイズデータのノイズスコア

図 6 はすべてのサンプルにおけるノイズスコアの分布を示す．閾値を設けていないが，検出されたノイズは右側に高いスコアで分布しており，配置場所と分布形態において他のサンプルと異なる性質を持っていることがわかった．

5. ShadowN vs Confident learning

5.1 使用データ

ShadowN と CL の比較に使用するデータは，Nematzadeh ら（2020）に使用した 6 つのベンチマークデータである（

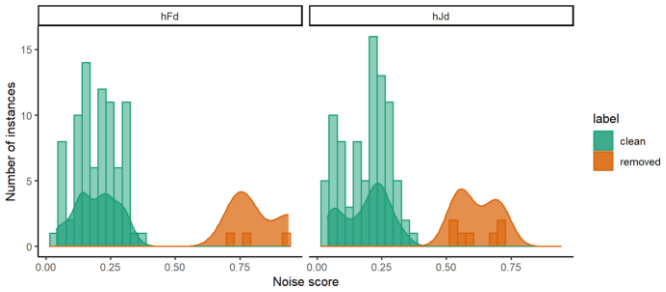


図 6　すべてのサンプルにおけるノイズスコアの分布

表 1）．これらのデータはいずれも実世界からとったデータであり，分類タスクとラベルノイズ検出によく使用される．すべてのデータのクラス数は 2 である．

表 1　使用データ

データ	#サンプル	#変数	クラス比	説明
Pima	768	8	0.54	糖尿病患者の身態特徴
Wisconsin	569	30	0.59	乳癌患者のデジタルイメージデータ
Liver	583	9	0.40	肝臓患者の記録
Parkinson	195	22	0.33	パーキンソン病患者の生体音声データ
Heart	299	12	0.47	心臓病患者の観察データ
Ionosphere	351	32	0.56	レーダーデータ

5.2 評価指標

適合率 Precision，再現率 Recall と F-score を評価指標とする．それぞれの計算は以下に示す．

$$\text{Precision} = \frac{\text{正しく検出されたノイズの数}}{\text{検出されたノイズの数}}$$

$$\text{Recall} = \frac{\text{正しく検出されたノイズの数}}{\text{実際にあるノイズの数}}$$

$$\text{F-score} = (1+\beta^2)\frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}}$$

式に示されているように，適合率とは，ノイズであると予測したもののうち，実際にノイズであったものの割合であり，正常データを誤って予測する確率を小さくすることを重視する際に使う指標である．再現率は，実際にノイズであるもののうち，正しくノイズと予測できたものの割合であり，ノイズを誤って予測する確率を小さくすることを重視する際に使う指標である．F-score は適合率の再現率の調和平均である．先行研究により，データの過剰なクレンジングはデータの元構造を崩壊させる一方，データの量という視点から学習精度向上を妨げるため，ラベルノイズ検出では適合率に重みをおくべきである．そこで，先行研究を参照し，βを 0.5 に設定する．

5.3 Confident learning の実装モデル

CL はモデルに依存して動くため，SVM（support vector machine），RF（random forest）と GTBoost（gredient tree boosting）を用いる．これらのモデルを選んだ理由は，数多くの研究に有効性が検証されたこと，幅広く使用されていることと，信頼性が高いライブラリが公開されて簡単に実装できることである．本研究では，Python の scikit-learn を使用する．また，すべてのモデルの性能を最大に発揮するため，データごとにチューニングを行う．

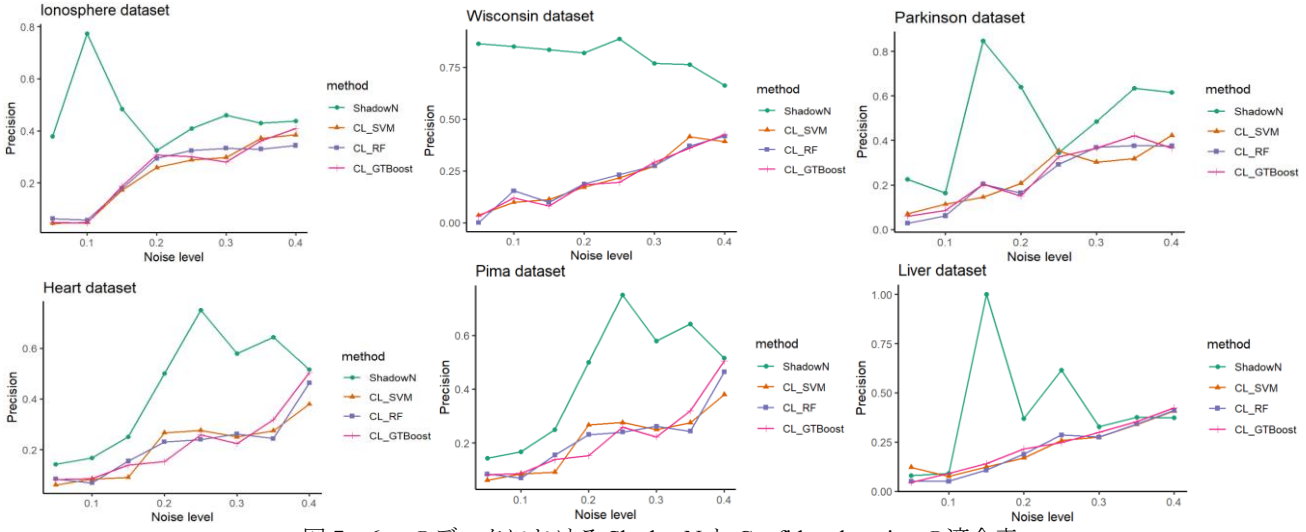


図 7　6 つのデータにおける ShadowN と Confident learning の適合率

5.4 結果

ノイズの割合を調整し，ShadowN と CL の適合率，再現率と F-score を比較した．ノイズ割合の設定について，先行研究では最大 50%まで設定したが，正常データとノイズが半分ずつ混じったデータに対しては，学習が極めて難しくなり，モデルの予測結果の信頼性が落ちる．また現実的にもこのようなデータは少ないと考えられる．さらに，Northcutt ら（2021）により，使用した 10 個のデータ（サンプル数 7,532~50,426,266）から平均的に 3.3%のノイズを検出した．以上により，本研究は，ノイズの存在は少数であることを仮定した上で，0.05 から 0.40 まで 0.05 刻みで 8 段階のノイズを注入した．

6 つのデータに対して ShadowN と CL の適合率を図 7 に示す．ノイズ割合の増加により，CL のノイズ検出における適合率が大きくなる傾向がある．一方で，ノイズが 10%以下のデータに対して，適合率は常に 0.2 以下であった．このように，CL はノイズが少数に存在するデータに対して適応しないことが示された．これに対して，ShadowN の性能とノイズ割合の変動とのあいだに一定の関係は見られなかった：ノイズが多くなるほど適合率は高くなるケース（Heart，Pima），ノイズが多くなるほど適合率は小さくなるケース（Ionosohere，Wisconsin），適合率がいったん高くなった後低下するケース（Parkinson，Liver）があった．これについては，ランダムに生成されたシャドウデータに影響を受けている可能性が考えられる．

適合率からみると，ShadowN は優位性を持っているが，データによっては，ShadowN のパフォーマンスが変動するため，ShadowN の安定性は CL より劣っていると言える．例えば，Wisconsin データにおいて， ShadowN の適合率は最小 0.66，最大 0.89 であった．一方で，Liver データにおいて，少数ノイズの検出に良い性能を示した ShadowN の適合率は，SVM を用いた際の CL より小さかった．このような差が出た原因を探るため，主成分分析で Wisconsin と Liver データの分布をプロットした（図 8）．Wisconsin データのクラスはある程度分けられており，Liver データのクラスはほぼ重なっている．ShadowN は重なりが大きいデータには適合しないことを示唆している．これは KNN からの影響である可能性が大きい．しかし一方で，重なりが大きいということから，元データの中に誤ったラベルが存在する可能性が考えられる．ノイズを注入するため，サンプルのラベルを真のラベルに入れ替えた可能性もある．この点に関して，更なる検討が必要である．

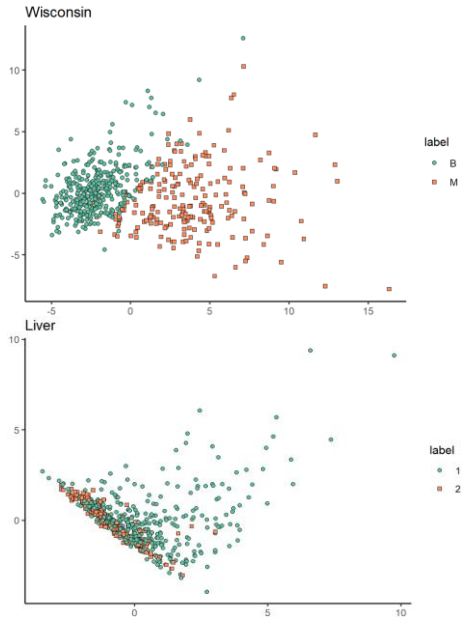


図 8　主成分分析による Wisconsin と Liver データ

6. まとめ

本研究では，ラベルノイズ検出方法 ShadowN を提案した．最新の方法である Confident learning と比較した結果，ノイズ割合が[0.05，0.40]の間になるデータには ShadowN はよりよい性能を示した．しかし，Confident learning はノイズ割合にかかわらずどのデータに対しても同じ傾向を示したことに対し，ShadowN はデータの性質に影響され，不安定である．ShadowN の安定性を改良することが今後の課題である．

参考文献

[1] Nematzadeh Z, Ibrahim R, Selamat A, “Improving class noise detection and classification performance: A new two-filter CNDC model”, Applied Soft Computing, 94, URL https://doi.org/10.1016/j.asoc.2020.106428 (2020).

[2] Northcutt CG, Athalye A, Mueller J, “Pervasive label errors in test sets destabilize machine learning benchmarks”, 35th Conference on Neural Information Processing Systems, Track on Datasets and Benchmarks, URL https://doi.org/10.48550/arXiv.2103.14749 (2021).