修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission: 01/11/2023 (MM/DD/YYYY)

専攻名（専門分野） Department	経営システム 工学専攻	氏 名 Name	宋　林鑫 Linxin Song	指 導 教 員 Advisor	後藤　正幸　　印 Seal
研究指導名 Research guidance	情報数理応用研究	学籍番号 Student ID number	CD 5221C026-3		
研究題目 Title	弱教師あり学習に基づくノイズデータクエリアルゴリズムの構築に関する一考察 A Study of Noisy Label Query Algorithm enhanced by Weak Supervision				

1. 研究背景と目的

機械学習の分野でノイズデータに対処するための一つのアプローチとして，能動学習（Active Learning）のプロセスを導入し，学習データセットからノイズラベルをクエリし，オラクル（Oracle, 人間レベルのアノテーションを提供できる主体）によってノイズデータのラベルを付け直す方法がある．そのモチベーションは，過半数のクリーンデータ（正解ラベルを持つ訓練データ）から学習したノイズ識別器はクリーンデータの特徴を記憶していることから，ノイズデータのクリーンラベルにも高い信頼度を付与できるであろうという点にある．しかし，従来研究では深層学習モデルをノイズ識別器として扱うことが多いが，深層学習モデルは大量なクリーンデータで事前学習させることができない場合，クエリアルゴリズムの性能がランダム選択とほぼ同等になってしまう．このような問題は「コールドスタート問題」と呼ばれ，能動学習において，人工的にデータにラベルを付与する予算が少ないときにしばしば発生する．

上記のコールドスタート問題に対する解決策の１つは，データの空間中の表現（Representation）に基づく能動学習手法である．これらの手法は，特徴変換のために他の関連データセットで事前に学習されたモデルを利用して変換された特徴間の距離に注目する．そしてその特徴を利用し，データセットからデータをクエリとする．しかし，データの表現に基づく手法には主に，(1) 事前に学習したモデルの性能に大きく依存するため，最適な変換特徴空間が得られない可能性がある，(2) 膨大なラベルなしデータに対しての計算量が大きい，という２つの問題点がある．上記の問題点により，データの表現に基づく手法は限られた量のデータしかクエリとすることができない．また，サンプリングされたデータはアンバランスになる可能性が高い．さらに，小規模のデータで学習する場合，最尤推定（MLE）の最適化プロセスはバイアスを生む可能性が高いという問題もある．このようなバイアスは，小さなサンプルのトレーニング済みモデルのエラー率を増加させてしまう．

本研究では，ラベルを付与するコストを低く抑えつつロバストなノイズ識別器を学習することを可能とした，弱教師あり学習に基づくノイズデータクエリアルゴリズムの構築を目的とする．従来研究では，分類タスクの実験において，弱教師あり学習は手動アノテーション付けなくても有望な結果を提供することが示されている．弱教師あり学習には，複数のノイズ付き教師ありソース（クラウドソーシングなど）とラベルなしデータセットを入力としてエンドモデル（End Model）を学習することでラベルを生成する２段階法がある．本研究ではまず，ノイズ識別器を事前学習するため，２段階法になる弱教師あり学習の手法：Adaptive Ranking based Sample Strategy (ARS2) を提案する．ARS2 は弱教師あり学習に基づくのモデル非依存フレームワークであり，以下の特徴を持つ．(1) ARS2 によってウォームアップされたモデルは，ノイズが多いデータに対して高い信頼性のある結果を与えることができる，(2) 実験結果により，与えられた高い信頼性のある結果は真のラベルである可能性が高い．これら２つの特徴により，本研究では最もアノテーションコストが低い弱教師ある学習手法 ARS2 を用いて，ノイズ識別器として扱うモデルを事前学習する．さらに，提案する ARS2 を用いたノイズデータクエリアルゴリズムを構築し，ラベル付与のコストとロバスト性の観点から有効な手法であることを示す．

本研究の主な貢献は，（1）弱教師あり学習により強化されたノイズデータクエリアルゴリズムを提案した，（2）文書分類の実データに基づく実験結果，既存のベースラインの性能を向上させることが実証したことの２点である．

2. 準備

2.1. 能動学習

能動学習の目標は，ラベルなしデータセットから最も費用対効果を持つ少数のデータを用いて，モデルのパフォーマンスを最大化することである．そのモチベーションは，現実の多くの場合，データにラベルを付与するコストが高い（医療データ，自動運転の画像データなど）ということにある．最も費用対効果の高いデータを探し出すため，能動学習は深層学習モデルの「不確実性」によってデータを選択する．例えば，２値分類の場合，不確実性に基づくクエリアルゴリズムは，正である事後確率が 0.5 に最も近いサンプルをクエリす

る．より一般的な不確実性サンプリングアルゴリズムでは，エントロピーを不確実性の尺度として利用する．

$$\{x^*_{ENT}\} = \text{top-}k \ -\sum_{i=0}^{C} P(y_i \mid \boldsymbol{x};\boldsymbol{\theta}) \log P(y_i \mid \boldsymbol{x};\boldsymbol{\theta}), \tag{1}$$

ここで，$\boldsymbol{x}$ は入力データ，$\boldsymbol{\theta}$ はモデルのパラメータ，$\{x^*_{ENT}\}$ はクエリされたデータ，C はクラス数，y_i はラベルである．

ただし，この様なクエリアルゴリズムでは，正確な不確実性を測定するため，事前に学習されたモデルが必要である．データ数が不足する場合では，不確実性に基づくクエリアルゴリズムはランダム選択と同等になってしまう．本研究では，弱教師あり学習を用いて，アノテーションコストがゼロのモデルを事前学習し，事前学習済みのモデルを使用して不正確な能動学習の推定を調整することにより，精度の高いノイズデータクエリアルゴリズムを提案する．

2.2. 弱教師あり学習

弱教師あり学習には，複数のノイズの多い教師ありデータソースとラベルのないデータセットを入力とし，エンドモデルを学習するための学習ラベルを生成する方法（2 段階法），または手動アノテーションなしでダウンストリームタスクの最終モデルを直接生成する方法（1 段階法）の２種類がある．２段階法の代表的な手法の１つは COSINE [1] である．COSINE は対照学習 (Contrastive Learning) の手法を用いて，同じクラスのデータのマージンを最小に，異なるクラスのデータのマージンを最大にし，加えてエラー伝播を防ぐためのソフトラベルを使用することで，弱教師あり学習の精度向上に効果を示した．１段階法についての代表的な手法の１つは Denoise [2] である．Denoise はアテンションネットワークを用いて，ラベリング関数に重要度に基づいて重みを付ける．そして重み付きラベリング関数のアノテーション結果を多数決の手法を用いて１つの学習ラベルに集約する．さらに集約された学習ラベルを用いてエンドモデルを訓練する．Denoise は弱教師あり学習タスクにおいても目立つ効果がある．しかし，上記の２つの手法はラベルを修正し直すことはできず，さらに不均衡データセットに対して性能が悪いという問題がある．

3. 提案手法

本節では，深層学習モデルの事前学習に用いる弱教師あり学習の手法である Adaptive Ranking based Sample Strategy（ARS2）とこれを用いたノイズデータクエリーアルゴリズムの２つの提案について，その詳細を示す．

3.1. ノーテーション

データセットを $\mathcal{D}$，データ数を N，各データを $\boldsymbol{x}_i \in \mathcal{X}$ とする（$i = 1, 2, ..., N$）．各データ $\boldsymbol{x}_i$ に対応するラベルを $y_i \in \mathcal{Y} = \{1, 2, ..., C\}$ とし，データセットの中にはノイズがあると想定する．能動学習のプロセスにより，本研究は深層学習モデル $f(\cdot;\boldsymbol{\theta}) : \mathcal{X} \to \Delta^{C}$[1] を R 回繰り返し学習する．ここで，第 r 回の学習済みモデルを $f^{(r)}(\cdot;\boldsymbol{\theta})$ で表す．

Algorithm 1: ARS2

Input: 弱ラベル付けされたトレーニングデータ $\mathcal{X}$; 仮ラベル $\tilde{y}$; 分類モデル $f^{(0)}(\cdot;\boldsymbol{\theta})$.

// ウォームアップ

for $t = 1, 2, \cdots, T$ **do**

1. $\mathcal{X}$ からミニバッチ $\mathcal{B}$ をサンプリング.
2. Early-stop までに，式 (2) を用いて θ を更新.

// サンプル選択により引き続き学習.

for $t = 1, 2, \cdots, T_s$ **do**

1. $x \in \mathcal{X}$ に対しての PMS を計算.
2. $\mathcal{X}$ から $\mathcal{Q}^{(t)} = \bigcup_{r_i \in \boldsymbol{r}} \underset{x \in \mathcal{X}_{r_i}}{\text{top-k}}\ s(x)$ をサンプリング.
3. $\mathcal{X}$ から $\mathcal{S}^{(t)} = \bigcup_{y_i \in \mathcal{Y}} \underset{x \in \mathcal{X}_{y_i}, s(x) > \xi}{\text{top-k}}\ s(x)$ をサンプリング.
4. $\mathcal{U}^{(t)} = \mathcal{Q}^{(t)} \cup \mathcal{S}^{(t)}$ を用いて θ を更新.

Output: 事前学習した $f^{(0)}(\cdot;\hat{\boldsymbol{\theta}})$ を出力.

本研究で用いた弱教師あり学習手法 ARS2 は２段階法である．この手法では，k 種類のヒューリスティックルール (Heuristic Rule) $\{r_i\}_{i \in \{1,...,k\}}$ を用い，各データに対しての弱ラベルを付与する．各ルール r_i は特定のラベル $y_{r_i} \in \mathcal{Y}$ に関連付けられており，ルール r_i の出力は $l_i \in \mathcal{Y} \bigcup \{-1\}$ である．ここで，ルールによりデータのラベルを判断できない場合は $l_i = -1$ とする．次に，統計的なラベルモデル (Label Model) を用い，ルールの出力 $\vec{l}$ を１つの仮ラベル (Pseudo Label)$\tilde{y} \in \mathcal{Y}$ に集約する．本研究は ARS2 と $\tilde{y}$ を用いて，$f^{(0)}(\cdot;\boldsymbol{\theta})$ を事前学習する．

3.2. ARS2 を用いたモデルの事前学習

データ数不足の問題を解決するため，本研究は弱教師あり学習を用いた Adaptive Ranking-based Sample Selection (ARS2) を提案し，能動学習の最初段階のモデル $f^{(0)}(\cdot;\boldsymbol{\theta})$ を事前学習する．ARS2 のアルゴリズムを Algorithm 1 に示す．まず，ラベルモデルで集約した仮ラベル $\tilde{y}_x$ を用いて分類モデル $f^{(0)}(\cdot;\boldsymbol{\theta})$ をウォームアップ (Warm-up) して，次の段階でノイズを識別できるノイズ判別機としてモデルを学習させる．次に，Class-wise Ranking (CR) および Rule-aware Ranking (RR) はサンプリングアルゴリズムによりサンプリングされたデータを用いて，ウォームアップされたモデルを継続学習 (Continual Learning) する．

ARS2 では，各学習データがクリーンかどうかを判断するために，ウォームアップ時に学習されたノイズ判別器を活用する．ただし，十分な記憶力を持つ深層学習モデルは，複雑なデータのパターンを「記憶」することで，学習データを過学習していまい，汎化性能が落ちる可能性が高い．モデルがノイズの多いデータで過学習することを防ぐために，early-stopping を導入し，以下の最適化問題を解く．

$$\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{x \in \mathcal{X}} \mathcal{L}(f^{(0)}(x;\boldsymbol{\theta}), \tilde{y}_x). \tag{2}$$

[1] Δ^C は C 次元の記号.

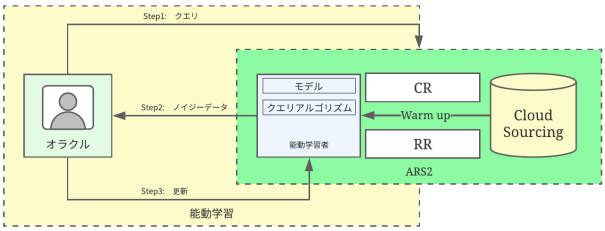

図 1: 提案手法の流れ

ここで，$\mathcal{L}$ は損失関数を表し，$\tilde{y}_x$ はラベルモデルによって集約された仮ラベルである．ARS2 では，多分類タスクに適した任意の損失関数を使用できる．

3.3. ノイズデータクエリーアルゴリズム

ARS2 による事前学習後，深層学習分類モデルは，クリーンラベルに高い信頼性を与えることができる．ノイズの多いデータの場合，事前学習済みのモデルは，入力データの特徴と前提知識に基づいてより正しい信頼度の分布を与えることができる．それは，大きなモデル学習では，いくつかのクリーンなサンプルに基づいて適切な特徴空間を学習し，さらに学習した特徴空間に基づいてノイズの多いデータの不確実性を出力できるためである．ただし，不確実性の出力で最も信頼できるクラスは，クリーンラベルになる可能性が高い．上記の分析に基づいて，最も信頼できる不確実性の出力に従って，以下のノイズデータクエリ手法を提案する．

$$Z^{(r)} = \left\{ (\boldsymbol{x}, y) \middle| y \neq \max_y (f_y^{(r)}(\boldsymbol{x}; \boldsymbol{\theta})) \right\} \bigcup Z^{(r-1)}, \quad (3)$$

ここで，$Z^{(r)}$ は r 回目でクエリされた学習セットであり，そのラベルはオラクルによって再修正される．ただし，第 0 回のモデルは ARS2 アルゴリズムの出力 $f^{(0)}(\cdot; \hat{\boldsymbol{\theta}})$ である．さらに，$Z^{(r)}$ を用いて，次の最適化問題を解くことで，$r+1$ 回目のためのモデルを更新する．

$$f^{(r+1)}(\cdot; \boldsymbol{\theta}) = f^{(r)}(Z^{(r)}; \boldsymbol{\theta}), \quad (4)$$

$$\text{where } \boldsymbol{\theta}^{(r+1)} = \min_{\boldsymbol{\theta}} \frac{1}{|Z^{(r)}|} \sum_{x \in Z^r} \mathcal{L}(f^{(r)}(x; \boldsymbol{\theta}), \tilde{y}_x).$$

提案手法の全体的な流れを図 1 に示す．

4. 評価実験

4.1. 実験設定

提案された方法の有効性を検証するため，ベンチマークデータセット AGNews 用いた評価実験を行う．AG News は4つのラベルを持つニュースデータセットである．データセット全体を，それぞれ 96,000，12,000，および 12,000 のデータでト学習，検証，およびテストセットに分ける．ARS2 の事前学習では，文献 [2] により提供される弱ラベルを使用する．パフォーマンスの向上を直感的に比較するため，ARS2 と同様のアンバランスなデータになるように，データセットを調整する．具体的には，AG News を 4 つの不均衡比率 $\rho \in \{1, 10, 20, 50\}$ を用いてアンバランス セットを作成した（ここで $\rho = \max_y \mathbb{P}(\boldsymbol{x}) / \min_y \mathbb{P}(\boldsymbol{x})$ である）．

能動学習では，比較的低コストで，ノイズの多いデータセットから最もノイズを持つデータの部分集合を探し出すことが目標である．上記の動機に基づいて，すべてのクエリステージでの抽出データ数を 100 に設定し，$R = 10$ 回繰り返す．これは，最終的に AG News に 1,000 件のデータ (データセット全体の 1.2%) をアクセスすることを意味する (コストは 1000 個のデータと同じ)．まだ，バックボーン分類モデルとして RoBERTa [3] を使用した．事前学習では，ARS2 ウォームアップ段階でノイズ識別機として RoBERTa を訓練するのではなく，ウォームアップされた MLP をノイズ識別機として RoBERTa の訓練バッチをサンプリングする．すべてのデータセットの評価指標として，テストセットの分類 F1 のマクロ平均スコアを使用した．PyTorch と WRENCH コードベース [4] を使用して実装した．

4.2. 実験結果

本研究は提案手法と 2 つの弱教師あり学習の従来手法 (COSINE, Denoise) と比較して提案手法の有用性を検証する．その結果を表 1 に示す．

表 1 により，提案手法を使用した ARS2 は，すべてのデータセットでパフォーマンスが向上していることがわかる．再修正されたラベルの数は，不均衡率が大きくなると低下する傾向を示す．その原因は，バランスの取れたデータセットで事前に学習されたモデルは，少数クラスのノイズの多いデータをより正確に判断できるためである．これは，モデルが圧倒的に不均衡な多数クラスのデータの影響を受けにくいことを示している．対照的に，アンバランスなデータセットに対して，事前学習モデルは少数クラスのノイズデータの均等な信頼分布を出力しており，少数派クラスのノイズデータに対する正しい判断の数が低下することがわかる．

また，不均衡率の増加に伴い，提案手法を使用した場合のパフォーマンス向上も顕著になる．これは，モデルがバランスデータセットで事前に十分に学習されているためと考えられる．不均衡なデータセットの場合，特徴空間と決定境界は十分に学習されていないため，修正されたクリーンなデータで訓練しながら簡単に変換することが可能である．

4.3. アブレーション研究

提案手法の各モジュールの貢献を示すため，本研究では提案手法の「ARS2+ラベル修正（ノイズデータクエリ)」，及び ARS2 だけ使用する「ARS2」と ARS2 でウォームアップして，ランダムでデータを抽出してラベルを修正する「ARS2+ラベル修正（ランダム)」の 3 手法を比較する．その結果を表 1 に示す．この結果より，ランダムでラベルを修正することで一定のパフォーマンス向上が可能であり，提案のノイズデータクエリ法（式（3）, 式（4)）を用いることで，さらにモデルのパフォーマンスを向上させることができることがわかる．

表 1: 提案手法を使用した場合と使用しない場合の ARS2 の比較.

	AG News($\rho = 1$)	AG News($\rho = 10$)	AG News($\rho = 20$)	AG News($\rho = 50$)
COSINE	85.4(0.4)	72.0(6.8)	82.2(0.3)	57.4(0.4)
Denoise	85.2(0.1)	53.7(11.1)	52.6(11.2)	47.1(18.8)
ARS2	88.2(0.3)	85.9(1.2)	84.4(3.5)	82.7(1.1)
ARS2 + ラベル修正 (ランダム)	89.7(1.4)	89.2(1.6)	86.4(2.4)	85.5(2.7)
ARS2 + ラベル修正 (ノイズデータクエリ)	**91.4(1.0)**	**90.9(0.7)**	**89.2(0.4)**	**88.4(1.2)**
ノイズデータクエリのうち修正したラベル	642	631	561	437

1. 括弧内の数値は標準偏差である.

表 2: AG News でクエリされたデータのサンプル ($\rho = 10$)

No.	サンプル	ノイズラベル	修正したラベル
1	China grabbed its first-ever men's Olympic gold in track and field on Friday, but it was a night of misery for Americans, who saw any chance of a fourth consecutive gold in basketball crushed.	1	2
2	Users of the Treo smartphone will get out-of-the-box compatibility with Microsoft's Exchange Server, thanks to a licensing deal between Milpitas, Calif.	3	4
3	Google unveiled a desktop search agent yesterday that lets users find any information stored on their computer, including visited Web pages, files and e-mail as well as instant messages.	3	4
4	A host of companies led by Microsoft (Quote, Chart) and Intel has revised and released a specification for making sure computer systems have a common way to communicate.	3	4

ラベル 1∼4 は，1. 政治，2. スポーツ，3. ビジネス，4. テクノロジーに対応してする.

4.4. 考察

ここでは，表 2 にリストされている，事前学習済みモデルによってクエリされたデータのサンプルの具体的事例について考察してみる．ケース 1 では，クラウドワーカーが「China」と「America」に従って「政治」ラベルを付与している．しかし，ケース No.1 の主な内容は，オリンピックの結果とそれに対応する市民の感情に関するものであり，政治ニュースとの関係性は低いと考えられる．その結果，ケース 1 をコンテンツに適した「スポーツ」ニュースとして再ラベルが付与している．ケース No.2，3，4 の場合，クラウドワーカーは，会社名の内容に従って，これらのニュースに「ビジネス」のラベルを付与している．しかし，ケース 2 は，スマートフォンユーザーと Microsoft のサービスとの互換性に関するものである．ケース 3 は，Google が提供する新しいデスクトップ検索エージェントのニュースに関するものである．ケース 4 は，Microsoft と Intel の間の新しい技術協力に関するものである．これらのケースは，主に企業の動きに関する内容であるが，主に技術に焦点を当てているため，テクノロジーに関連しているように思われる (例：新しいサービス，新しいエージェント，技術協力)．上記の考え方に基づいて，ケース 2，3，4 を「テクノロジー」に分類する方がより合理的と判断する.

5. 結論と今後の計画

本研究では，ロバストな能動学習のためのノイズデータクエリアルゴリズムを提案した．MLE には小データセットに対してのバイアスがあり，以前の方法では深層学習の設定でバイアスを処理できない．また，以前の不確実性に基づく能動学習手法は，データが少ない場合に性能が低下してしまう．これらの制限に対処するため，弱教師あり学習手法の ARS2 でモデルを事前学習して，ノイズデータを再修正し，修正されたデータを用いてモデルを学習する手法を提案した．実オープンデータセットである AG News を使用した実験により，提案された方法がノイズラベル付けされたデータを発見することができ，最先端の方法と比較して F1-macro スコアが向上することを示した．また，ケーススタディを通じて，以前のアノテーターとオラクルの観点からラベルを再修正する方法を説明し，アノテーションに関する合理的な詳細を提案した.

今後の課題としては，(1) より効果的な深層学習モデルの事前学習手法の提案，(2) 別のタスクのデータセットを用いた提案手法の有効性の検証，(3) よりソフト的なクエリアルゴリズムの構成の 3 点が挙げられる.

参考文献

[1] Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. *arXiv preprint arXiv:2010.07835*, 2020.

[2] Wendi Ren, Yinghao Li, Hanting Su, David Kartchner, Cassie Mitchell, and Chao Zhang. Denoising multi-source weak supervision for neural text classification. *arXiv preprint arXiv:2010.04582*, 2020.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[4] Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. Wrench: A comprehensive benchmark for weak supervision. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.