補助事業番号 2023M-342
補 助 事 業 名 2023年度 クラウドソーシングにおける表現学習によるコスト削減方法 補助事業
補助事業者名 山梨大学 李吉屹

1 研究の概要

本研究は、クラウドソーシングサービスを利用したデータへの高精度ラベル付与タスクにおいて、トレードオフの関係にあるコスト削減と品質向上を同時に目指す。表現学習によるオブジェクトに対する各ワーカーのラベルを予測し、補完したバランスのとれたラベル行列をラベル推定モデルに利用する方法を提案する。提案方法は2つの自己教師付き信号に対する2つの分類タスクをマルチタスクモデルで学習し、スパースなラベルに潜在する情報を効率的に学習する。不均衡なラベル数とオブジェクトの曖昧さを考慮した損失関数も提案する。この課題に関する情報を調査した。モデルを考案した。提案モデルを開発・検証・評価・改良した。論文を執筆した。提案方法は、クラウドラベルから効果的に表現を学習することができ、特にクラウドラベルがスパースの場合に、正解ラベル推定のパフォーマンスを向上させることができる。 また、2023年から、大規模言語モデル（LLM）が注目を集めている。LLMがラベル付与タスクにおいてクラウドソーシングより良いかどうか検証した。クラウドラベルとLLMラベルの両方を正解ラベル推定モデルに使用することで、特にクラウドラベルがスパースな場合に、推定された正解ラベルの品質が向上し、クラウドラベルやLLMラベルのみを用いた場合よりも高くなることを検証した。

2 研究の目的と背景

深層学習に代表される機械学習手法が容易に利用できるようになったことを背景に、教師付き機械学習の精度は、訓練データの質と量に依存する。しかし、人手による大量のデータ作成は多大な労力とコストを要する。近年、インターネットを介し不特定多数のワーカーにラベル付け作業を依頼するクラウドソーシングサービスの利用が増加している。しかし、ワーカーは必ずしもその作業に精通していないことが多いため、ラベルの質が担保されていないのが現状である。このような背景から、データの品質管理はクラウドソーシングにおける重要な課題の一つである。

代表的な解決策は、収集したラベルに冗長性を持たせることである。各オブジェクトに対して複数のワーカーにラベル付与を割り当ててもらい、そのラベルを統合することで、正解ラベルを推定することができるのである。ただし、オブジェクト数が多い場合、コストを考慮すると、各ワーカーは不特定多数のオブジェクトを持つ小さなサブセットに対してのみラベルを提供し、各ワーカーのラベル数は不均衡になり、オブジェクトとワーカーのラベル行列はスパースになる。この問題は、正解ラベル推定モデルが十分に学習されていないことに起因していると考えられる。

本研究の目的は、クラウドソーシングサービスを利用したテキストや画像などデータへの高精度ラベル付与タスクにおいて、表現学習によるラベル付与が必要なオブジェクトとワーカーの特質に着目することにより、スパースなクラウドラベルから高品質なラベルを推定し、サービス利用時のコストを抑えることが可能な方法を提案する。本研究は機械学習を対象としたラベル付与にお

いてトレードオフの関係にあるコスト削減と品質向上を同時に目指す点が挑戦的であり、独自性がある。

3　研究内容

（1）スパースクラウドラベルにおける表現学習によるコスト削減と品質向上方法に関する研究
(https://drive.google.com/file/d/1_Ak9_EKj1Oy6ucrhswAQRNbgW112DXIk/view)

クラウドソーシングサービスを利用したデータへの高精度ラベル付与タスクにおいて、トレードオフの関係にあるコスト削減と品質向上を同時に目指す。表現学習によるオブジェクトに対する各ワーカーのラベルを予測し、補完したバランスのとれたラベル行列をラベル推定モデルに利用する方法を提案する（図１）。提案方法は2つの自己教師付き信号に対する2つの分類タスクをマルチタスクモデルで学習し、スパースなラベルに潜在する情報を効率的に学習する。不均衡なラベル数とオブジェクトの曖昧さを考慮した損失関数も提案する。この課題に関する情報を調査した。モデルを考案した。

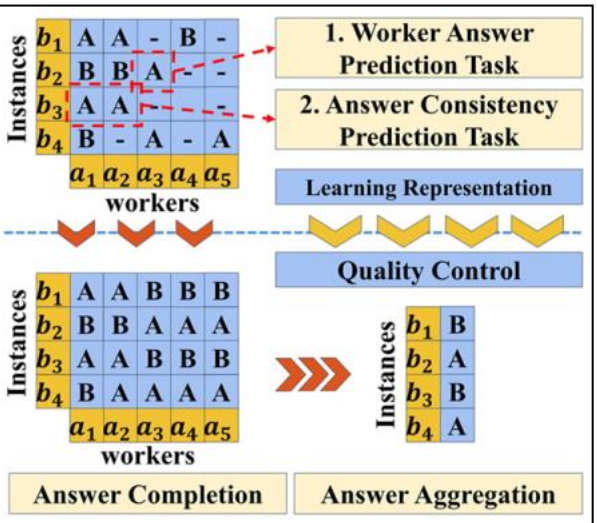

図１　表現学習によるクラウドラベル品質向上の提案方法のフレームワーク

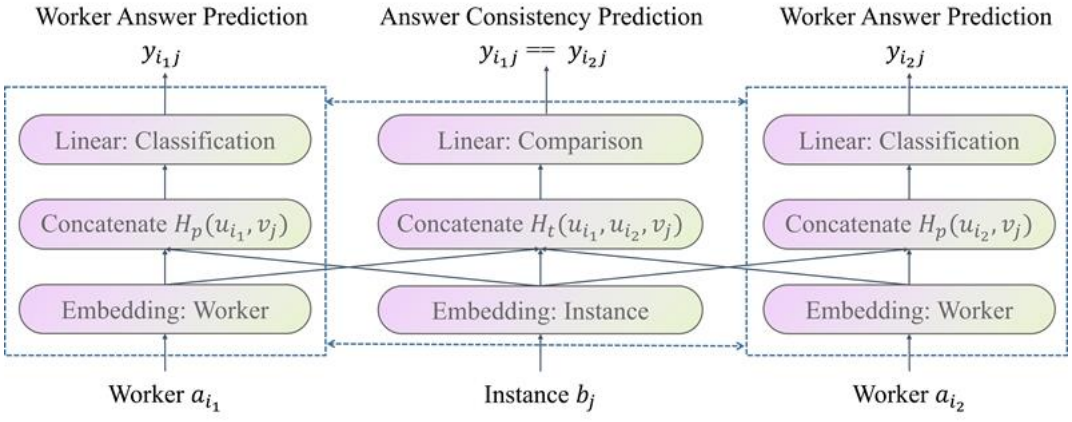

図２　提案された自己教師あり表現深層学習モデル

提案モデルを開発・検証・評価・改良した。論文を執筆した。提案方法は、クラウドラベルから効果的に表現を学習することができ、特にクラウドラベルがスパースの場合に、正解ラベル推定のパフォーマンスを向上させることができる。

また、2023年から、大規模言語モデル（LLM）が注目を集めている。LLMがラベル付与タスクにおいてクラウドソーシングより良いかどうか検証した。既存のクラウドソーシングデータセットのうち、

どのデータセットを比較研究に利用できるかを調査し、ベンチマークを作成した。個々のクラウドラベルとLLMラベルの品質を比較し、正解ラベル推定モデルを用いて、推定された正解ラベルに対する評価を行う。クラウドとLLMのハイブリッド正解ラベル推定モデルを提案し、開発・検証・評価した。論文を執筆した。クラウドラベルとLLMラベルの両方を正解ラベル推定モデルに使用することで、特にクラウドラベルがスパースな場合に、推定された正解ラベルの品質が向上し、クラウドラベルやLLMラベルのみを用いた場合よりも高くなることを確認しました。

4　本研究が実社会にどう活かされるか―展望

深層学習などの教師付き機械学習において学習データの作成に直接貢献することから、産業界における多様な分野での人工知能技術の実用化と進展が期待できる。クラウドソーシングにおいてラベルのコスト削減と品質向上による信頼できる技術が社会に実装され、安心して信頼性の高いAIを利用できる人間中心のAI社会実現に貢献する。産業に対して、信頼性が高く効率的な人工知能を導入することは有益である。人間に対して、自分の知識や能力を合うクラウドソーシングタスクを探し完成できるし、誰もが個性や能力を発揮できる環境の整備に貢献できる。

5　教歴・研究歴の流れにおける今回研究の位置づけ

研究代表者李吉屹（Jiyi Li）は、クラウドソーシングデータの信頼性と統合に関する研究を精力的に行っており、トップや主要な国際学会で発表した。さまざまな種類のラベルに対応するクラウドラベル統合課題を取り込んだ。最近、ノイズの多いクラウドソーシングデータから直接深層学習モデルを学習することに取り組んだ。本研究課題は、本研究で必要とされるクラウドソーシングと深層学習における研究代表者の研究成果を基に提案されたものである。

今後の課題として、提案方法を拡張することで、コンテンツ情報を活用した正解ラベル推定方法、低品質ワーカーをブロックし、高能力ワーカーを選出する方法、大規模な不均衡データに対するコスト削減可能なラベル付け方法を提案することができる。ラベル付与で生じる問題点は、自然言語処理などのメディア処理にも還元することができることから、本研究の学術的意義は極めて大きい。

6　本研究にかかわる知財・発表論文等

国際会議論文（査読付）

[1]. Jiyi Li, ″Learning Representations for Sparse Crowd Answers″, Proceedings of the 30th International Conference on Neural Information Processing (**ICONIP 2023**), pp. 468–480, Nov. 2023.

[2]. Jiyi Li, ″A Comparative Study on Annotation Quality of Crowdsourcing and LLM via Label Aggregation″, Proceedings of the 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing (**ICASSP 2024**), pp. 6525–6529, Apr. 2024.

7　補助事業に係る成果物

（1）補助事業により作成したもの

論文URL：

[1]. Learning Representations for Sparse Crowd Answers

https://doi.org/10.1007/978-981-99-8076-5_34

https://dl.acm.org/doi/10.1007/978-981-99-8076-5_34

https://link.springer.com/chapter/10.1007/978-981-99-8076-5_34

[2]. A Comparative Study on Annotation Quality of Crowdsourcing and LLM via Label Aggregation

https://doi.org/10.1109/ICASSP48485.2024.10447803

https://ieeexplore.ieee.org/document/10447803

https://arxiv.org/abs/2401.09760

8　事業内容についての問い合わせ先

所属機関名：　山梨大学工学部（ヤマナシダイガク コウガクブ）

住　　所：　〒400-8511

山梨県甲府市武田4丁目3-11 山梨大学工学部コンピュータ理工学科

担 当 者：　准教授　李　吉屹（リ ジイ）

担 当 部 署：　コンピュータ理工学科（コンピュータリコウガッカ）

Ｅ－ｍａｉｌ：　jyli@yamanashi.ac.jp

ＵＲＬ：　http://www.sites.google.com/site/jiyilisite/