様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 ２８ 日現在

機関番号：１３５０１
研究種目：若手研究
研究期間：2019～2021
課題番号：１９Ｋ２０２７７

研究課題名（和文）クラウドソーシングによるコスト削減のため機械学習法

研究課題名（英文）Machine Learning Methods for Cost Reduction in Label Collection by Crowdsourcing

研究代表者
李　吉屹（Li, Jiyi）

山梨大学・大学院総合研究部・助教

研究者番号：３０７２６６６７

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究は，テキストや画像などメディアを対象とした大規模データへのクラウドソーシングによる高精度ラベル付与タスクにおいて，サービス利用時コストを抑えることが可能な機械学習法を明らかにすることである．クラウドソーシングによって収集したデータの精製を行い，ラベル付与の曖昧さを排除する手法を提案した．事例とワーカーを選出することによって，データ品質を向上させる方法を提案した．多様なメディアにおいてモデルを構築するために，カテゴリーラベルの回答統合方法をシーケンスなどの多様なデータ型に対応できるように拡張した．人工知能に関するトップ国際会議IJCAI，WWW，SIGIR，MMを含む論文８編を発表した．

研究成果の学術的意義や社会的意義
本研究は，大規模データと多数のラベルを対象としたラベル付与においてトレードオフ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．テキストや画像など実用レベルで利用可能な機械学習モデルを提案することであり，ペアワイズラベル及びシーケンスラベルへの拡張にも挑戦する．ラベル付与で生じる問題点は，機械学習及び自然言語処理など人工知能分野にも還元することができることから，学術的意義は極めて大きい．近年脚光を浴びている深層学習などの教師付き機械学習において本質的な問題である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．

研究成果の概要（英文）：The objective of this study is to propose machine learning methods that can reduce the cost of using the crowdsourcing service in the task of accurately annotating large-scale data for various media processing, such as text and images. We proposed methods for disambiguating label assignment by refining data collected through crowdsourcing. We proposed methods to improve data quality by selecting instances and workers. In order to build models in various media, by incorporating the content of the instances, we extended the methods of answer aggregation with categorical labels so that it can handle diverse data types such as sequences. We have published 8 papers at international conferences including the top international conferences on artificial intelligence such as IJCAI, WWW, SIGIR, and MM.

研究分野：クラウドソーシング、データマイニング、自然言語処理

キーワード：クラウドソーシング　ラベル付与　コスト削減　機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

深層学習に代表される先進的な機械学習手法とそれを支えるテキストや画像，音声などのビッグデータが容易に利用できるようになったことを背景に，人工知能技術が脚光を浴びている．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世界中の不特定多数のワーカーにラベル付け作業を依頼し対価を支払うクラウドソーシングサービスの利用が急激に増加している．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，対価に見合うラベル付け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り，これまでに数多くの研究が行われてきた．初期の研究では，複数のワーカーが付与したラベルの多数決を用いる方法が試みられていた．しかし，ワーカーの多くは非専門家であることから，単純な多数決では質を向上させることは困難である．そこで，一人のワーカーが複数のデータに対してラベル付け作業を行っているという点に注目し，データ間の情報を共有することによりラベル付けの質を向上させる試みがなされている．

ラベル付けに関する高い品質を保つため，ワーカーに注目した研究も多数行われている．例えば，能力が高いワーカーをできるだけ早期に発見するために，ワーカーの正解率を推定するモデルが提案されている．しかし，これまでの研究の多くは，クラウドソーシングサービスを利用して得られたデータに対して，如何にして高品質なデータを取得するかという問題に焦点をあて，確率モデルや機械学習手法によりデータを取得する方法がとられていた．さらに，それらの多くは人工的に生成したデータを用いて検証を行っているため，スパース性やスケーラビリティの課題が依然として残されている．

２．研究の目的

本研究の核心をなす「問い」は，「機械学習の本質的な問題である訓練データの作成において，如何にして高品質なデータを低コストで作成するか」である．本研究の目的は，クラウドソーシングを利用した大規模データへの高精度ラベル付与タスクにおいて，ラベル付与が必要なデータとワーカーの特質に着目することにより，サービス利用時のコストを極力抑えることが可能な機械学習法を提案することである．

本研究の目的は，クラウドソーシングを利用した大規模データへの高精度ラベル付与タスクにおいて，ラベル付与が必要なデータとワーカーの特質に着目することにより，サービス利用時のコストを極力抑えることが可能な機械学習法を提案することである．本研究の学術的独自性は，(i) 予めデータの精製を行い，ラベル付与の曖昧さを極力排除する点，及び(ii) ラベル付与の対象となるデータの選別と専門家ワーカーの選出を一つの処理とみなし，これらを有機的に統合することにより高品質なラベル付与を行う点の 2 点に集約できる．大規模データを対象としたラベル付け作業において，ラベルを付与する際に生じる曖昧さや既存のラベルに属さない事例の存在は，ラベル付きデータの品質を下げる大きな要因の一つとなっている．本研究はこれらに対し，事例が持つラベル数を推定すると同時に，既存のラベルに属さない事例を例外として自動的に認識・排除する新しい枠組みを提案することにより品質の低下を防ぐ点に，オリジナリティがある．

ラベル付けのコストを抑え，高品質なラベル付与を行うためには，大量のデータから機械学習の精度に貢献するデータを選別し，そのデータのみにラベル付与作業を依頼すればよい．このための一つの手法として従来能動学習が広く用いられてきた．本研究ではこれを拡張し，精度に貢献し，かつラベル付けが困難なデータを抽出すると同時に，能力の高いワーカーにこのラベル付け作業を依頼する手法を開発する点が独創的である．ラベル付与が必要なデータの抽出，あるいはタスクに依存して優秀なワーカーを選出する手法についてはこれまでに多くの手法が提案されている．本研究がこれらと根本的に異なる点は，両者を有機的に統合する点であり，これにより最小限のコストで高品質なラベル付与が可能となる．ラベルを付与するデータの量にも注目し，データの分布を推定した結果を用い，ワーカーにラベル付けが必要な事例を提示する点にもオリジナリティがある．

本研究のゴールは，テキストや画像など実用レベルで利用可能な機械学習モデルを提案することである．そこで，ペアワイズラベル，及びシーケンスラベルへの拡張にも挑戦する．同時にラベル付与で生じる問題点は，現在の機械学習理論，及び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．

３．研究の方法
本研究は，テキストや画像など各種メディア処理を対象とした大規模データへのクラウドソーシングによる高精度ラベル付与タスクにおいて，サービス利用時のコストを抑えることが可能な機械学習法を明らかにすることである．本研究は 3 つの課題から成る．
[課題 1: データ精製]
クラウドソーシングによって収集したデータの精製を行い，ラベル付与の曖昧さを排除する手法を開発した．多腕バンディットと困惑度に基づく探索法によるラベル収集の予算コスト削減法を提案した．部分的なランクデータによって感情分布を持つ多様で信頼性の高く感情ラベルを効率的に収集するための感情画像ラベル付与技術を提案した．クラウドトリプレットラベルに対する回答統合法を提案した；データを公開した．
[課題 2: 事例選択とワーカー選出]
事例とワーカーを選出することによって，データ品質を向上させる方法を提案した．ワーカーのパフォーマンスを制約条件及び正則化として利用する回答統合法を提案した．クラウドから収集した人間のテキストアイデアにおける，アイデア内容と優先度スコアの関係を構築に基づく優先度統合方法を提案した．
[課題 3: ペアワイズラベル，及びシーケンスラベルへの拡張]
多様なメディアにおいて機械学習モデルを構築するために，事例の内容に注目し，カテゴリーラベルの回答統合方法をシーケンスラベルに対応できるように拡張した．クラウド単語シーケンスに対する回答統合法を提案した；データを公開した．ハイブリッド信頼性とハイブリッドテキスト表現に基づく方法を提案した．
本研究は，大規模データと多数のラベル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．近年脚光を浴びている深層学習などの教師付き機械学習において本質的な問題である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．

４．研究成果
人工知能に関するトップ国際会議 IJCAI，WWW，SIGIR，MM を含む，査読ある国際会議論文８編を発表した，うち第１著者６編，国内会議論文１編を発表した．２つデータセットを公開した．クラウドソーシングの研究に関する本研究のチームの論文や公開されているデータセットを下記の URL に公開している．https://github.com/garfieldpigljy/ljycrowd
(1) [AnnoNLP 2019] クラウドソーシングにおいて，クラウドソーシングされたワードシーケンスの最初のデータセットを作成と公開した，ワードシーケンスの正解を作成のため作業者信頼性を考慮した回答統合法を提案した．データセットを公開した．
データセット: Crowdsourced Word Sequence Aggregation 2019
https://github.com/garfieldpigljy/CrowdWSA2019

(2) [ICONIP 2019] クラウドソーシングにおいて，多腕バンディットと困惑度に基づく探索法によるラベル収集の予算コスト削減法を提案した，各事例の必要なラベル数を推定して，ラベル収集のコストを削減と同時に，ラベルの有用性は保証できる．

(3) [IJCAI 2020] 作業者のパフォーマンスを正解回答を推測するためのグローバルな制約条件として利用することを提案した．この制約条件を既存の統計的統合法と組み合わせた正則化として利用することも提案した．実験では，制約条件が単独で使用された場合に正解を推定力を持つだけでなく，正則化として使用された場合には既存の統合法を後押しすることを示した．

(4) [SIGIR 2020] 多様な質を持つ複数のクラウドソーシングによるテキストを統合するために，事例に対する回答の局所的な信頼性と，作業者に対するデータセットに全局的な信頼性というハイブリッドな信頼性情報を取り込むことができる統合方法を提案した．局所的な信頼性については，テキストの類似性をテキストの埋め込みと単語列というハイブリッド表現から取り込む．実データを用いた実験により，ベースラインよりも優れていることを示した．

(5) [HCOMP 2020] クラウドから集められたアイデアの集合が与えられたとき，アイデアの価値に関する多様な潜在的評価基準を考慮したクラウドの評価者による選好比較に基づいて，優先順位をつけるために，少なくとも 1 つの潜在的評価基準の観点から最も優れたアイデアのサブセットを得るための方法を提案した．実データを用いた実験結果により，提案手法が複数の視点からアイデアに効果的に優先順位をつけることができることを示した．

(6) [MM 2020] Game With a Purpose (GWAP)の概念に基づき，感情分布を持つ多様で信頼性の高く感情ラベルを効率的に収集するための新しい感情画像ラベル付与技術 AffectI を提案した．部分的なランキングデータを収集システムを開発し，データを統合する方法を提案した．選択，推定，インセンティブという 3 つの新しい機構を備えている．実験の結果，既存の手法と比較し

て，より多様で信頼性の高いラベルを収集できるという点で優れていることを示した．

(7) [ICONIP 2021] 人間のアイデアやデザインなどのオブジェクトの整理と解析において重要なのは，その類似性である．人間は相対的な判断が得意であるため，類似性の比較はトリプレット（二重ペアワイズ）で行われる．クラウドソーシングにおいて，トリプレットラベル（二重ペアワイズラベル）の正解を作成するため，作業者の能力と対象物の難易度を考慮した回答統合法を提案した．2 つの新しい実データセットを構築と公開した．提案モデルの候補から自動的に最適なモデルを探索する．実験により，提案方法の有効性が検証された．
データセット: Crowdsourced Triplet Similarity Comparisons 2021
https://github.com/garfieldpigljy/CrowdTSC2021

(8) [WWW 2022] クラウドから収集した人間のテキストアイデアにおける，社会的な意思決定のために優先順位をつける解決策を提案するために，アイデア内容と優先度スコアの関係を構築に基づく優先度統合方法を提案した．一般的な既存の優先度統合方法は，ペアごとの優先度ラベルを利用するしかない．アイデアのテキスト内容のような文脈情報と優先度スコアの外部関係または内部関係を構築することにより，同質的な設定と異質な設定の両方に対する方法を提案した．実クラウドデータを用いた実験により，提案方法は，特に利用可能な優先度ラベルの数が少ない場合に，集団の選好を推定するためのベースラインよりも優れた統合結果を生成できることが示された．

＜引用文献＞
[AnnoNLP 2019]. Jiyi Li, Fumiyo Fukumoto, "A Dataset of Crowdsourced Word Sequences: Collections and Answer Aggregation for Ground Truth Creation", Proceeding of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019) Workshop on Aggregating and analysing crowdsourced annotations for NLP (AnnoNLP 2019), pp. 24-28, Hong Kong, Nov. 2019.

[ICONIP 2019]. Jiyi Li, "Budget Cost Reduction for Label Collection with Confusability based Exploration", Proceeding of the 26th International Conference on Neural Information Processing of the Asia-Pacific Neural Network Society (ICONIP 2019), pp.231-241, Sydney, Dec. 2019.

[IJCAI 2020]. Jiyi Li, Yasushi Kawase, Yukino Baba, Hisashi Kashima, "Performance as a Constraint: An Improved Wisdom of Crowds Using Performance Regularization", Proceeding of the 29th International Joint Conference on Artificial Intelligence (IJCAI 2020), pp. 1534-1541, virtual event, Jul. 2020.

[SIGIR 2020]. Jiyi Li, "Crowdsourced Text Sequence Aggregation based on Hybrid Reliability and Representation", Proceeding of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020), pp. 1761-1764, virtual event, Jul. 2020.

[HCOMP 2020]. Yukino Baba, Jiyi Li, Hisashi Kashima, "CrowDEA: Multi-view Idea Prioritization with Crowds", Proceeding of the eighth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2020), Vol.8, No.1, pp.23-32, virtual event, Oct. 2020.

[MM 2020]. Xingkun Zuo, Jiyi Li, Qili Zhou, Jianjun Li, Xiaoyang Mao, "AffectI: A Game for Diverse, Reliable, and Efficient Affective Image Annotation", Proceeding of the 28th ACM International Conference on Multimedia (MM 2020), pp.529-537, virtual event, Oct.2020.

[ICONIP 2021]. Jiyi Li, Lucas Ryo Endo and Hisashi Kashima, "Label Aggregation for Crowdsourced Triplet Similarity Comparisons", Proceedings of the 28th International Conference on Neural Information Processing (ICONIP 2021), pp. 176-185, virtual event, Dec. 2021.

[WWW 2022]. Jiyi Li, "Context-based Collective Preference Aggregation for Prioritizing Crowd Opinions in Social Decision-making", Proceedings of the ACM Web Conference 2022 (WWW 2022), pp. 2657-2667, Lyon, Apr. 2022.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　8件）

１．発表者名 Jiyi Li, Yasushi Kawase, Yukino Baba, Hisashi Kashima
２．発表標題 Performance as a Constraint: An Improved Wisdom of Crowds Using Performance Regularization
３．学会等名 Proceeding of the 29th International Joint Conference on Artificial Intelligence (IJCAI 2020)（国際学会）
４．発表年 2020年

１．発表者名 Jiyi Li
２．発表標題 Crowdsourced Text Sequence Aggregation based on Hybrid Reliability and Representation
３．学会等名 Proceeding of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)（国際学会）
４．発表年 2020年

１．発表者名 Xingkun Zuo, Jiyi Li, Qili Zhou, Jianjun Li, Xiaoyang Mao
２．発表標題 AffectI: A Game for Diverse, Reliable, and Efficient Affective Image Annotation
３．学会等名 Proceeding of the 28th ACM International Conference on Multimedia (MM 2020)（国際学会）
４．発表年 2020年

１．発表者名 Yukino Baba, Jiyi Li, Hisashi Kashima
２．発表標題 CrowDEA: Multi-view Idea Prioritization with Crowds
３．学会等名 Proceeding of the eighth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2020)（国際学会）
４．発表年 2020年

１．発表者名
左幸坤，李吉屹，茅暁陽

２．発表標題
画像におけるゲームによる多様・信頼的・効率的な感情アノテーション

３．学会等名
第13回データ工学と情報マネジメントに関するフォーラム（DEIM2021）

４．発表年
2021年

１．発表者名
Jiyi Li, Fumiyo Fukumoto

２．発表標題
A Dataset of Crowdsourced Word Sequences: Collections and Answer Aggregation for Ground Truth Creation

３．学会等名
Workshop on Aggregating and analysing crowdsourced annotations for NLP (AnnoNLP 2019, conjunction with EMNLP-IJCNLP 2019)（国際学会）

４．発表年
2019年

１．発表者名
Jiyi Li

２．発表標題
Budget Cost Reduction for Label Collection with Confusability based Exploration

３．学会等名
the 26th International Conference on Neural Information Processing of the Asia-Pacific Neural Network Society (ICONIP 2019)（国際学会）

４．発表年
2019年

１．発表者名
Jiyi Li, Lucas Ryo Endo and Hisashi Kashima

２．発表標題
Label Aggregation for Crowdsourced Triplet Similarity Comparisons

３．学会等名
Proceedings of the 28th International Conference on Neural Information Processing (ICONIP 2021)（国際学会）

４．発表年
2021年

１．発表者名
Jiyi Li

２．発表標題
Context-based Collective Preference Aggregation for Prioritizing Crowd Opinions in Social Decision-making

３．学会等名
Proceedings of the ACM Web Conference 2022 (WWW 2022)（国際学会）

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Crowdsourced Word Sequence Aggregation 2019
https://github.com/garfieldpigljy/CrowdWSA2019

クラウドソーシングによるトリプレットの類似性比較データセット(CrowdTSC2021)
https://github.com/garfieldpigljy/CrowdTSC2021

クラウドソーシングの研究に関する本研究のチームの論文や公開されているデータセット
https://github.com/garfieldpigljy/ljycrowd

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関