

教師あり機械学習のアノテーションにおける

クラウドソーシングと機械学習の比較

杉村 太一[†] 目良 侃太郎[‡] 小坂 隆浩[‡]同志社大学理工学部[†] 同志社大学院理工学研究科[‡]

1. はじめに

近年、機械学習が著しく進歩しており、科学や産業において大きな影響を持つ技術となっている。教師あり機械学習を新しい分野へ応用する際に最初に必要となる作業が、アノテーションである[1]。アノテーションとは、教師あり機械学習にモデルを学習する際に必要となる学習データを作成することであり、例えば、犬の画像に「犬」、「Tom is tall and he is kind.」というテキストに「he=Tom」となる。

アノテーションで用いる手法は、クラウドソーシングのみを用いる、専門家のみを用いる、機械のみを用いる、人(クラウドソーシングまたは専門家)と機械を組み合わせる、などの方法が考えられる。本研究では、クラウドソーシングのみを用いる方法と、機械のみを用いる方法を用いた場合の、教師あり機械学習のアノテーションにおける精度と費用を比較し、アノテーションにおける方法選択の一助となることを目的とする。

2. 教師あり機械学習

教師あり機械学習は学習データ(ラベル付きのデータ)からクラス分類モデルを学習する学習フェーズと、学習済みのクラス分類モデルにラベル

なしデータを入力しラベルを出力(予測)する予測フェーズに分けられる。教師ありの機械学習の学習フェーズと予想フェーズに必要なデータと流れについて図1に示す。本研究では、学習フェーズにおいて学習に利用するラベル付きデータを作成することに焦点を当てる。

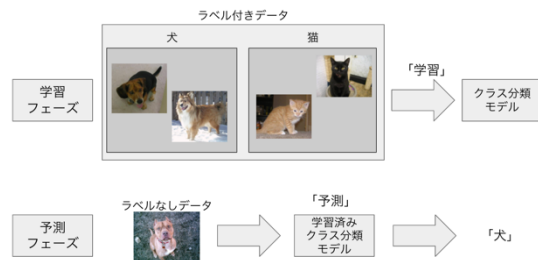


図1 機械学習の学習フェーズと予想フェーズに必要なデータと流れ

3. 評価実験

アノテーションにクラウドソーシングを用いる場合について実験を行う。クラウドソーシングを依頼するツールとしては、Amazon Mechanical Turk を用いる[2]。アノテーションを行うデータセットは、iNaturalist の 180 枚の画像で、1つのタスクを「1枚の画像を Plant/Insect/Bird/Reptile/Fungus/Amphibian/None of the above のいずれかに分類する」作業とする。1つのタスクを5人のワーカーに割り当て、1つのタスクにおける報酬を0.05ドルとした。クラウドソーシングの精度と費用は1つのタスクを何人のワーカーに割り当てる冗長度に依存するため、冗長度を変更した際

A Comparison of Crowdsourcing and Machine Learning in Supervised Machine Learning Annotation.

Taichi Sugimura[†] Kantaro Mera[‡] Takahiro Koita[‡]

[†] School of Engineering, Doshisha University

[‡] Graduate School of Engineering, Doshisha University

の精度と費用を、機械学習を同じデータセットに適用した際の精度と比較する[3]. 最終的な精度は、5人のワーカーの回答を多数決で集計したものと正解ラベルの一致度で算出した.

4. 結果と考察

実験では 180 個中 179 個のタスクが完了された. 実験で測定したクラウドソーシングの精度と費用, 先行研究の機械学習のベンチマークでの精度を表 1 に示す.

表 1 クラウドソーシングと機械学習の比較

	精度(%)	費用(\$)
クラウドソーシング	96.64	\$54
機械学習	81.20	—

表 1 から, アノテーションにおける精度については, クラウドソーシングの方が機械学習より 15.40%高くなっている. また実験で 1 つのタスクを 5 人のワーカーに割り当てているため, 1 つのタスクにあたり 0~5 個の誤答が含まれる可能性がある. 1 タスクにおける誤答数とそれに対応するタスクの数を表 2 に示す.

表 2 1 タスクあたりの誤答数

1 タスクにおける 誤答数	5	4	3	2	1	0
タスク数	0	3	3	7	25	141

表 2 から, 1 タスクあたりの誤答数を参考にして, 冗長度を変化させた時の精度の変化を推測すると, 1 タスクを 1 人にのみ割り当てた場合は, 179 個中最大で 25 個誤答する可能性があり, その場合は, 精度が 86.03%となる. これは冗長度を最低, つまり 1 人のワーカーのみの回答を参考にした場合の精度である. これは比較対象である機械学習のベンチマークより精度が高い. また, その際の

クラウドソーシングにかかる費用は, \$10.8 である. 以上より, アノテーションを行う際の方法としては, クラウドソーシングを用いた方法の方が, 機械学習を用いた方法より精度を高くすることが可能である. ただしより多くの費用を要する.

5. おわりに

本研究では, 教師あり機械学習のアノテーションにおいてクラウドソーシングを用いた方法と機械学習を用いた方法の精度と費用を比較した. クラウドソーシングの精度と費用を算出するために実験を行い, 機械学習の精度は先行研究の結果と比較した. 結果として, 冗長性を持たせない場合でもクラウドソーシングを用いた方が精度が高くなることを示した. 今後はクラウドソーシングを用いた方法において, データセットの数を増す, 冗長度を高める, 報酬を下げた場合の精度の変化を測定する予定である.

参考文献

- [1] S.B. KOTSIANTIS. “Supervised Machine Learning: A Review of Classification Techniques”. Emerging Artificial Intelligence Applications in Computer Engineering I. Maglogiannis et al. (Eds.) IOS Press, 2007.
- [2] Amazon Mechanical Turk.
<https://www.mturk.com/>
- [3] Hugo Touvron, Alexandre Sablayrolles, Matthijs Douze, Matthieu Cord and Hervé Jégou. “Graft: Learning fine-grained image representations with coarse labels”, arXiv:2011.12982v1 [cs.CV], Cornell University, 2020.