The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

人間ニューラルネットワーク

Crowd Neural Network

坂田 雄亮 *1　馬場 雪乃 *2　鹿島 久嗣 *1*3
Yusuke Sakata　Yukino Baba　Hisashi Kashima

*1京都大学 Kyoto University
*2筑波大学 University of Tsukuba
*3理化学研究所 革新知能統合研究センター RIKEN Center for AIP

Feature values are indispensable for almost all machine learning methods to make predictions; however, the definitions of the features themselves are sometimes too abstract so that they cannot be extracted automatically. Human-in-the-loop machine learning is a promising solution to such cases where humans extract feature values for machine learning models. In this research, we use crowdsourcing for feature extraction and consider a problem to aggregate the feature values given by crowdsourcing workers to improve the accuracy of machine learning classifiers. We propose a novel neural network model called *Crowd Neural Network* that learns both the capabilities of human feature extractors and the weights of a neural network classifier by applying the idea of the convolution neural network to feature aggregation. Our experiments using four datasets show the proposed method outperforms the baseline method using an existing unsupervised aggregation method in some datasets.

1. はじめに

機械学習技術の近年の発展は著しく、様々な分野において機械学習の応用が行われているが、いずれにおいてもまずは学習対象の情報を計算機可読な特徴量によって表現する必要がある。しかし、その特徴量の抽象性が高い場合には、その機械的抽出もまた困難であり、従って機械学習の適用も困難となる。一方で、機械的処理では抽出困難な特徴であっても、人間にとっては比較的容易に抽出可能な特徴量も存在する。このような場合、まずは人間が特徴量を抽出し、これを用いて機械学習を行うといった、いわゆる人間参加型 (human-in-the-loop) のパイプラインが有効であろう。従来はこのようなパイプラインの実現には、人的リソースのコストが大きかったが、近年のクラウドソーシングプラットフォームの台頭により人的リソースの大量かつ機械的な獲得が可能となり、人間を特徴抽出機構と見なし機械学習アルゴリズムに組み込む事が容易となった。

機械と人間の協調が必要となる問題の例としては、絵画の作者を推定する問題や、動画中の人物が作り笑いしているか否かを見分ける問題などが考えられる。これらは機械にとっても専門知識のない人間にとっても比較的困難な問題であろう。一方、絵画が明るいイメージであるかとか、動画内の人物が肩を震わせて笑っているかなどといった、最終的な判別に寄与すると期待されるやや抽象的な特徴は、依然として機械にはその判断が難しい反面、人間ならば比較的簡単に判断できる。このような抽象的な特徴量抽出をクラウドソーシングによって行うことで前述のパイプラインを実現できると期待できる。

しかし、クラウドソーシングにおいては、ワーカの能力ややる気のばらつきがあるため、結果として得られる成果物の品質にばらつきが生じる。そこで頑健性の確保の為に複数のワーカから回答を得て統合を行うことで信頼度の高い結果を得るという手法が広く用いられている。1 つのタスクについて複数のワーカからラベリングが行われている時、最も簡単な意見統合手法としては多数決や平均を取ることが挙げられる。

しかし単純に多数決を取ることにはいくつかの問題がある。あるタスクについて正解と不正解のラベルがある状況を仮定すると正解を見極める能力を持つワーカが少数の場合、単純に多数決を行なってしまうと不正解のラベルが選択されてしまう。そのためワーカの能力を考慮して意見統合を行う手法が広く研究されている。Dawid と Skene によって提案されたワーカの感度と特異度を用いて意見統合を行う手法はこの分野の先駆け的手法である [Dawid 79]。

また別の問題としてスパムワーカの存在がある。スパムワーカとはクラウドソーシングなどにおいて報酬のみを目的にタスクを意図されない形でこなすワーカなどを指し、スパムワーカの回答は大抵の場合実験の意図に則さないデータ、即ちノイズデータとなる。もちろんプラットフォーム側でも対策は行なっているものの完全な排除は難しくデータにスパムワーカによるものが混入することは考慮されなければならない。1 つのタスクに参加するワーカの数が多ければ多数決でもノイズは無視できるが、ワーカの数が少ない場合は結果に大きく影響が出てしまう。この場合もワーカの能力を考慮に入れることでスパムワーカの意見の影響を抑える事が可能であり結果的にノイズの除去が可能となる。

一般に広く知られる意見統合手法は統合したラベルを得る事が目的となるため、意見統合の際に正解を参考にできない一種の教師無し学習と捉える事ができる。一方、統合したラベルを学習に用いる本研究の設定の場合ラベル統合の明示的な正解ラベルは存在しないものの、ラベルを特徴量として使用した分類器の出力に正解が存在するので間接的に教師データを持つと言える。従って一度にワーカの能力と分類器を学習する手法は既存の意見統合手法と分類器の学習を組み合わせたものより優れた精度を実現しうると考えられる。

本研究ではクラウドソーシングを用いた機械学習の一手段として複数のクラウドワーカによってデータの特徴抽出を行った際、ワーカの能力まで学習対象として分類器の学習を行うことでより精度の高い分類器を作成する手法を提案する。ここで意見統合自体はワーカの能力を係数とする重み付き足し合わせで表現される。複数のクラウドワーカより得られる特徴ラベルの重み付き足し合わせを 1 次元の畳み込みを用いて実現し、畳み込みニューラルネットワークを用いて各ワーカの能力と各ノードとの重みを同時に学習する。

連絡先: 坂田　雄亮，京都大学大学院情報学研究科知能情報学専攻，yskskt75@ml.ist.i.kyoto-u.ac.jp

提案手法の分類精度及び適応範囲を確認するため、絵画の作者の識別、ニュース記事の公共性の識別、ホテルのレビュー文の真偽の識別、動画上の人物の作り笑いの識別の 4 つのデータセットに対しての実験を行ない、提案手法が既存手法に優る例を示した。

2. 関連研究

2.1 機械学習におけるクラウドソーシングの活用

クラウドソーシングの台頭によって不特定多数の人間に安価に作業を依頼することが出来る様になり機械学習にもクラウドソーシングを活用する研究が行われてきた。機械学習においてはまず学習に用いる教師データの収集の目的でクラウドソーシングが用いられた [Ipeirotis 10]。しかしクラウドソーシングを用いて得られるデータは品質のばらつきが大きい為、何らかの処理によって品質管理を行う必要がある。品質管理手法の一例として、一つのタスクに関して複数のワーカから結果を収集し統合する事で頑健な結果を得る手法が挙げられる。最も単純な意見の統合は多数決だが、より信頼度の高い結果を得る為に様々な手法が提案されている。

Dawid と Skene が提案したワーカの能力をワーカが正しく回答する確率によってモデル化する手法もその一つである [Dawid 79]。ワーカの能力、正解ラベルの二つが最終的に推定したいものであり、片方が判明していればもう片方が推定できる事を利用して、EM アルゴリズムによってこれらを交互に推定する。推定値の初期値としては仮の正解ラベルとして多数決によるものが使用される事が多い。

2.2 クラウドソーシングによる特徴ラベル生成

あるタスクの答えを得るためにクラウドソーシングを用いる場合、タスクの答えそのものを募集するのが最もシンプルな手法となる。しかしタスクが抽象的でかつある程度の専門性を必要とする課題である時、上記の手法には専門的知識を持つワーカが必要となる。そこで専門的知識を持たないワーカでも答えられる簡単な、しかし機械的に抽出する事が難しいような質問を用意しその回答を特徴ラベルとして教師付き機械学習に用いる手法が提案されている。これは本実験の問題設定に最も近い関連研究である。上述したタスクの例として画像を見て野鳥の種類を特定するタスクが考えられる。この場合、「お腹の色」や「くちばしの形」などの質問によって専門的知識を持たないワーカからも特徴ラベルを得る事ができる [Branson 10]。

2.3 クラウドソーシングによる特徴ラベル定義

機械学習の際にどの様な特徴量を定義するかは予測精度に大きく影響を及ぼすので、技術者は適切な特徴量の設定を行いたいが、しばしば適切な特徴量の設計には専門的知識や多大な労力が必要とされる事がある [Domingos 12]。そこで特徴ラベル生成に用いる質問そのものもクラウドソーシングを用いて作成する手法も提案されている。Flock と呼ばれるこのシステムはクラウドワーカに正例と負例を比べさせてその差異を記述させることで特徴の定義を行う [Cheng 15]。特徴ラベルの収集もまたクラウドソーシングによって行われる。より精度上昇に貢献しうる特徴定義を適当的に獲得する方法として AdaFlock が提案されている [Takahama 18]。AdaFlock ではブースティングの考え方を用いて、現在の分類器で誤分類されているデータをうまく分類出来るような特徴定義を逐次的に得る。

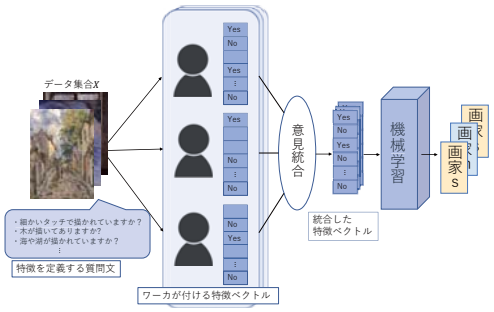

図 1: ラベリングタスクのイメージ図。例として Paintings データセットを用いた。

3. 問題設定

はじめに予測モデルの定式化を行う。本問題では、二値分類問題を基盤とする。X をサンプルの集合、$Y = \{-1, +1\}$ をラベルの集合とする。N 個のラベル付きサンプルからなる訓練データセット $D = \left\{(x_i, y_i) \in X \times Y\right\}_{i=1}^{N}$ を入力として、分類器 $H : X \to Y$ を獲得する事が目的となる。ここで、入力 X に関して特徴ベクトルは与えられておらず、X に対してクラウドワーカから特徴ラベルを収集する必要がある。結果として特徴ベクトルとしては i 番目のデータの k 番目の特徴に j 番目のワーカがつけた特徴ラベル X_{kj}^{i} が得られる。ただし各ワーカは全ての特徴をラベリングする必要はなく、欠損が存在していても構わない。また前提としてテストデータセットの特徴ラベリングに参加したワーカは全て訓練データセットの特徴ラベリングに参加している。即ち訓練データの学習を行うことで全ワーカの能力が学習できる状況を想定している。

なお、特徴は 'Yes'、'No' で答えられる質問にワーカが答える形でラベリングされるものとする。質問文は事前に準備したものを用いる。質問文の作成の詳細は Takahama ら [Takahama 18] に譲る。絵画のデータセットを例にあげると、X としては絵画の画像データの集合しか与えられていない。よって絵画の各画像データに対してクラウドワーカに「空が大きき書かれていますか？」や「輪郭がぼやけていますか？」といった質問により特徴ラベルを生成する。図 1 に特徴ラベリングタスクのイメージ図を示す。

4. Crowd Neural Network

特徴ラベリングタスクを経て各データについて複数のワーカから特徴ラベルが得られた。この複数のワーカからなるラベルを良く統合するためのワーカの能力の推定と、統合されたラベルを用いる分類器の学習を同時に行うニューラルネットワークである Crowd Neural Network を提案する。以下 Crowd Neural Network を Crowd NN と表記する。

4.1 ワーカ畳み込み層

提案手法では分類器の学習と同時にワーカの特徴抽出能力の推定を行う。ワーカの能力の定義にはいくつかの手法が存在するがここでは重み $\alpha = \alpha_1, \ldots, \alpha_J$ によって定義し、データの特徴ベクトルとして各ワーカの付けたラベルを重み付きで足し合わせたものを用いる事を考える。これは確率的モデルではなく、能力の高いワーカの意見を重視する単純な意見統合モデルとなっている。

以上の構造を実装するためにクラウドワーカの回答統合を行うワーカ畳み込み層を提案する。これは一般に用いられる畳

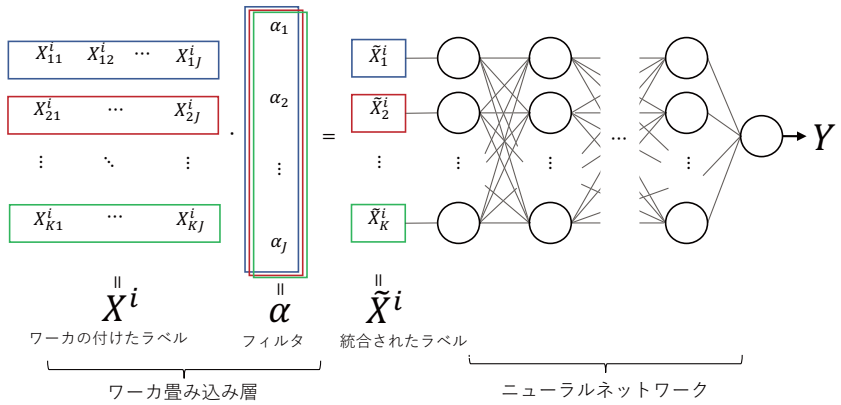

図 2: Crowd NN のイメージ図。同色の行とフィルタ α を掛け合わせて $\tilde{X}^i$ を計算しニューラルネットワークの入力とする。

み込みニューラルネットワークの畳み込み層を応用したものである。以下にワーカ畳み込み層内での処理を述べる。

まず特徴ラベルの重み付き足し合わせを畳み込みによって実現する構造について述べる。畳み込みの 1 ステップはフィルタという重みを共有する概念によってフィルタの範囲内の入力データをフィルタの重みに応じて足し合わせる事である。よって i 番目のデータの k 番目の特徴についてクラウドワーカがつけたラベル $X^i_{k1}, \ldots, X^i_{kJ}$ に対して、ワーカの能力に対応する 1 次元のサイズ J のフィルタとして α を用いて畳み込みを行う事でワーカの能力に応じた重み付き足し合わせを表現でき、統合した特徴ラベル $\tilde{X}^i_k$ を得られる。欠損している特徴ラベル、つまり該当ワーカが回答していないラベルについては 0 で埋められているため、ラベルの統合に影響しない。同様に全ての特徴に畳み込みを適用する事で統合した特徴ベクトルである $\tilde{X}^i$ が得られる。ここでフィルタの重みは全データ、全特徴で共有される事に注意されたい。

4.2 Crowd NN の学習

前述 のワーカ畳み込み層によって統合された特徴ベクトル $\tilde{X}^i$ は後段のニューラルネットワークの入力となる。特徴ベクトル $\tilde{X}^i$ に基づいて分類を行い発生した誤差の逆伝搬によってフィルタ、及び各ノードの重みを学習する。ニューラルネットワークの次元数、層数、及び活性化関数は後述するチューニングを行って決定する。

I をデータの総数、J をワーカの総数、K を特徴の総数とし i, j, k をそれぞれのインデックスとするとワーカ畳み込み層について順伝播の式は以下の様に表せる。

$$\tilde{X}^i_k = \sum_{j=1}^{J} X^i_{kj}\alpha_j$$

ここで得られた $\tilde{X}^i_k$ を上述したニューラルネットワークの順伝搬に適用する。一方逆伝播の式は誤差関数を E で表すと

$$\begin{aligned}\frac{\partial E}{\partial \alpha_j} &= \sum_{k=1}^{K} \frac{\partial E}{\partial \tilde{X}^i_k}\frac{\partial \tilde{X}^i_k}{\partial \alpha_j} \\ &= \sum_{k=1}^{K} \frac{\partial E}{\partial \tilde{X}^i_k} X^i_{kj}\end{aligned}$$

以上の様に表す事が出来る。ここで $X^i_{kj} = 0$ の時、すなわちデータ i の特徴ラベリングタスク k についてワーカが応答をしていない時はデータ i の特徴 k の誤差にワーカの能力が影響されない事がわかる。またバックプロパゲーションの誤差

$$\delta^i_k = \frac{\partial E}{\partial \tilde{X}^i_k}$$

はニューラルネットワークの一層目のレイヤーより逆伝播されている為、上式によってパラメータを更新する事が出来る。最後に Crowd NN のイメージ図を図 2 に示す。

4.3 特徴ラベルの取り扱い

特徴ラベル X^i_{kj} は欠損を含んでいても良く、欠損値は 0 で穴埋めされる。ここで質問文の返答として本質的意味合いを持つのはポジティブな反応である 'Yes' でありネガティブな反応である 'No' は大きな影響を持たないと考える事も出来る。その為ワーカの返答の入力ラベルの符号化を $\{1, 0, -1\}$ と $\{1, 0\}$ の二種類考える事とする。$\{1, 0\}$ の符号化を用いる場合、欠損データも 0 で埋めているのでネガティブな反応である 'No' は応答しなかった事と同義に扱われる事となる。

5. 評価実験

5.1 特徴ラベルデータセット

4 種類の特徴抽出タスクを対象にして、それぞれ 1 つの特徴量に対して特徴ラベリングを行うワーカ数を 10 人とし特徴ラベルを収集した。これらの特徴抽出タスク（使用する画像等のデータ、質問文）は、Takahama ら [Takahama 18] が設計したものと同じである。4 種類のタスクは、絵画の著者識別の Paintings、動画の人物の自然な笑顔と作り笑いの識別の Smiles 、ホテルに関する偽造レビューの識別の Reviews、記事がトップページに乗るほど重要か否かを識別する Articles である 。全てのデータセットは共通して正例、負例共に 200 件ずつ合計 400 件。各タスクごとの特徴量は 100 種とした。

5.2 ベースライン

提案手法と比較するベースラインとして以下の 2 つの意見統合手法を用いる：

1. 多数決: データ毎の特徴量をワーカの付けたラベルの単純な多数決によって得たもの。

2. Dawid&Skene (D&S): ワーカの感度と特異度ならびに特徴ラベルを EM アルゴリズムにより推定し特徴量とした

表 1: テストデータに対する精度の比較表。10 回の交差検証による精度の平均値と標準偏差を示す。また各データセットで最も精度が良かったものを太字で示す。

	Accuracy			
Dataset	Crowd NN	Pos. Crowd NN	多数決	D&S
Paintings	0.765 ±0.051	**0.773** ±0.039	0.732 ±0.065	0.762 ±0.066
Smiles	0.785 ±0.081	0.765 ±0.060	0.617 ±0.101	**0.787** ±0.073
Reviews	0.645 ±0.072	0.650 ±0.074	0.520 ±0, 064	**0.700** ±0.064
Articles	0.645 ±0.045	**0.710** ±0.087	0.550 ±0.031	0.660 ±0.055

もの。条件を揃えるために訓練データのみを用いてワーカの能力を推定する。

特徴ラベルの正解は存在しないため、統合したラベルを入力としたニューラルネットワークの精度で比較する。ニューラルネットワークのハイパーパラメータは提案手法と同様のチューニングを行う。

5.3 評価方法

各モデルの評価はハイパーパラメータのチューニングも含めた 10 分割交差検証を用いて行った。softmax の値に基づく精度の平均と標準偏差を評価値とした。ハイパーパラメータ及びその探索範囲として、各層のサイズを 100 から 300 まで 50 刻み、層数を 2 から 7 まで 1 刻み、活性化関数を *relu* または *leaky_relu* から選択、エポック数を 20 から 100 まで 10 刻み、バッチサイズを 20 から 200 まで 20 刻みとした。

5.4 実験結果

結果を表 1 に示す。なお、以降 $\{1, 0, -1\}$ の符号化を用いた提案法を Crowd NN、$\{1, 0\}$ の符号化を用いた提案法を Positive Crowd NN と呼ぶ。全てのデータセットを通じて Crowd NN は多数決より優れた精度となった。一方、D&S と比べて Painting データセットでは僅かに上回ったが Smiles データセットでは僅かに下回り、Articles データセット、Reviews データセットでは大きく下回る結果となった。この結果から Crowd NN は単純な多数決より良く意見統合が出来ている事が確認出来る。一方、D&S とは比較して必ずしも常に高い精度ではない結果となった。Positive Crowd NN は全てのデータセットを通じて多数決より優れた結果となった。Paintings、Articles データセットにおいては Crowd NN、D&S を上回る結果となった。Reviews データセットについて Crowd NN を僅かに上回ったが Dawid&Skene を下回り、Smiles データセットにおいて Crowd NN、D&S の両方に劣る結果となった。

Crowd NN が必ずしも Dawid&Skene よりも高い精度を得られなかった理由について考察する。まず挙げられるのは過学習の可能性であろう。今回の実験で用いた 4 つのデータセットはそれぞれ 400 件程度と比較的小規模である。既存手法に比べて、Crowd NN が学習するべきパラメータの数は多く、従って過学習が起きている可能性が指摘できる。

$\{1, 0\}$ の符号化を用いると $\{1, 0, -1\}$ の符号化を用いた場合に比べて精度が上がったデータセットでは、不要な特徴を省く事で学習が容易となり結果として過学習の軽減に繋がったと考えられる。ただしネガティブなラベルである 'No' が比較的価値をもつデータセットにおいてはこの情報が失われるため精度が下がったものと思われる。

6. 結論

本研究ではクラウドソーシングを利用して生成された特徴ラベルを用いた機械学習の精度向上を目指して、特に特徴ラベルの意見統合に焦点を当てた手法としてクラウドワーカの能力をフィルターに用いた畳み込みニューラルネットワークを提案した。この提案手法の性能を測るため 4 つのデータセットを用いて実験を行い提案手法の長所と短所を確認した。

今後の研究の展望としては予測精度への貢献性の低い特徴量の削減が考えられる。貢献性の低い特徴量は分類器の学習自体にはそもそも大きく影響することはなく、ワーカの能力推定に対してはノイズになっていると考えられる。通常の意見統合手法の場合は統合ラベルが目的であるが、統合ラベルを特徴量に機械学習を行う今回の設定の場合は統合ラベルは目的でないため精度予測への貢献性が低いラベルを無視する事が出来る。

データ数を増やす事で過学習を回避する事も考えられる。特徴生成にクラウドソーシングを用いているためコストとのトレードオフとなってしまうが、データ数の増加によって提案手法のより正確な性能を確認出来る事が期待される。

今回の実験では訓練データから特徴生成に参加した全ワーカの能力を推定する事が出来るという前提としているが、実際にはテストデータの特徴生成に新規ワーカが参加する可能性は無視できない。その様な事前に能力推定が出来ないワーカの能力の取り扱いについても検討の余地がある。

参考文献

[Branson 10] Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., and Belongie, S.: Visual Recognition with Humans in the Loop, in *Proceedings of the 11th European Conference on Computer Vision*, pp. 438–451 (2010)

[Cheng 15] Cheng, J. and Bernstein, M. S.: Flock: Hybrid Crowd-Machine Learning Classifiers, in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 600–611 (2015)

[Dawid 79] Dawid, A. P. and Skene, A. M.: Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, No. 1, pp. 20–28 (1979)

[Domingos 12] Domingos, P.: A Few Useful Things to Know About Machine Learning, *Communications of the ACM*, Vol. 55, No. 10, pp. 78–87 (2012)

[Ipeirotis 10] Ipeirotis, P. G.: Analyzing the Amazon Mechanical Turk Marketplace, *ACM XRDS*, Vol. 17, No. 2, pp. 16–21 (2010)

[Takahama 18] Takahama, R., Baba, Y., Shimizu, N., Fujita, S., and Kashima, H.: AdaFlock: Adaptive Feature Discovery for Human-in-the-loop Predictive Modeling, in *Proceedings of the 32nd Conference on Artificial Intelligence* (2018)