DEIM2021 I24-3

クラウドソーシングによる訓練データセット構築における最適な冗長度の検証

清水　綾女†　　若林　啓††

† 筑波大学　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †shimizu.ayame.sw@alumni.tsukuba.ac.jp, ††kwkaba@slis.tsukuba.ac.jp

あらまし　教師あり機械学習では，クラウドソーシングを用いた訓練データの作成が広く行われている．しかし，クラウドワーカの回答は品質が保証されないため，データの集約には工夫が必要とされる．一般的な対策として複数人から同じタスクに対する回答を得る冗長化が行われているが，クラウドソーシングにかけられる予算の大きさに上限があると考えると，冗長化は回答の精度を向上させる一方で訓練データの量を減少させる．本研究では，冗長度と訓練データ数の積を予算と定義し，予算の大きさ，ワーカの回答精度，学習器にとってのタスクの難易度の３変数から，クラウドソーシングの効用を最大化させる冗長度の検証をシミュレーションデータを用いた実験で行った．

キーワード　クラウドソーシング，冗長化，教師あり機械学習

1 はじめに

機械学習によるタスク解決は自然言語処理，画像認識など様々な分野で用いられているが，訓練のために大量の訓練データを用意する金銭的，時間的コストは未だ大きな負担となっている．そこで，コスト削減の対策としてインターネットを通して不特定多数に仕事を依頼するクラウドソーシングサービスを用いた訓練データ作成が広く行われている．

クラウドソーシングを用いることで，専門家にアノテーションデータ作成を依頼する場合より時間的，金銭的コストを削減できる．しかし一方で，不特定多数のクラウドワーカによる回答は精度の担保がされないため，利用にあたっては工夫が必要とされる．最も一般的な対策として同じタスクを複数のクラウドワーカに対して割り振る冗長化があり，冗長化したワーカの回答に対する効果的な集約法については多数の研究がある [1] [2] [3] [4] [5]．しかし，多くの場合クラウドソーシングにかけられる予算の大きさには上限がある．この場合，冗長化は訓練データの精度を向上させる一方で訓練データの量を減少させる．高い予測精度の学習器を構築するという目標において，クラウドソーシングを冗長化させることによって得られる効用を最大化させる手法 [4] [5] についての研究は盛んに行われている．しかし，訓練データの質を高める反面，訓練データの量を減少させるというデメリットを持つ冗長化を行う意義があるのかについては明らかになっていない．訓練データの量が学習器の性能向上において重要であることに加えて，ある程度のノイズが含まれる訓練データを用いても学習器の構築が可能である [6] [2] ことが知られており，訓練データの冗長化が必ずしも学習器の構築に効果的であるとはいえない．

クラウドソーシングを用いた訓練データ作成について，最適な冗長度を検証した従来の研究 [2] では，画像分類タスクにおける実験で冗長化しない（すなわち，冗長度が 1）とき最も優位となることが示された．しかし，従来の研究 [2] ではタスクの難易度や予算規模の影響を考慮していない．そのため，一般の知見が明らかになっておらず，従来の研究 [2] では冗長化の意義の有無は明らかになっていない．

本研究では，クラウドソーシングにおける冗長化が，訓練データの質を高める一方で訓練データの量を減少させるものであるという事実に着目し，クラウドソーシングを用いた訓練データ作成において最も効果的な冗長度の検証を行う．実験は予算の大きさ，クラウドワーカの回答精度，学習器にとってのタスク難易度を変数とし，シミュレーションデータを用いて行う．ここでは同じタスクを割り振る人数を冗長度とし，冗長度とデータ数の積を予算として定義する．クラウドワーカの回答集約では，多数決を用いる方法と，ワーカの回答をモデル化して真のラベルを推定する手法 [1] の２手法による検討を行う．冗長度ごとに集約した訓練データで学習器を構築し，それぞれの学習器の予測性能によって各条件における最適な冗長度を決定する．さらに，訓練データセットのデータ数が学習器構築に与える影響，および冗長度が訓練データセットの精度に与える影響をシミュレーションデータによる実験によって明らかにする．これらの実験からの冗長度と学習器構築の関係について考察し，クラウドソーシングを用いた機械学習器構築において冗長化に意義があるのかどうかを明らかにする．

2 先行研究

クラウドソーシングによる訓練データ作成における既存研究は，2 つの問題設定に大別される．１つはクラウドワーカの回答から真のラベルを推定するという問題設定であり，もう 1 つはクラウドソーシングで得たラベルを訓練データとして識別器を構成するという問題設定である．

1 つ目の問題設定に対してはワーカの能力をモデル化し，真のラベル推定を行う Dawid ら [1] の手法を拡張したものが大多

数を占める．Dawid ら [1] は EM アルゴリズムを用いて，ワーカの能力推定と真のラベル推定を交互に行う手法を考案した．Dawid ら [1] の手法を拡張した研究は数多くあり，特に固有表現抽出に応用させたものとしては Simpson ら [4] の研究や，Nguyen ら [5] の研究がある．固有表現抽出とは与えられた文章から，人名や地名などの固有名詞や時間表現や日付などの数値表現を抽出するタスクである．Dawid ら [1] の手法は医師の診断の統合のために考案されたため，タスク同士は独立しているものとしてモデル設定がされているが，固有表現抽出では与えられた文章中の単語同士は依存関係を持つと考えられる．そのため，Simpson ら [4] は前後のタスクとの依存関係を表現できるように Dawid ら [1] の考案したワーカの回答能力を表すモデルを拡張した．さらに，Nguyen ら [5] はクラウドソーシングにおいて，1 人のワーカから得られる回答データは僅かであることが多いということに着目し，ワーカあたりの回答数が少ない場合でもワーカの能力推定モデルを構築できるように Dawid ら [1] の手法を拡張した．

クラウドソーシングで得たラベルを訓練データとして識別器を構成するという問題設定は，機械学習の発展において特に重要である．Khetan ら [2] は学習器の予測結果をワーカの能力推定に用いることで，冗長化をせずともワーカの能力推定ができるように Dawid ら [1] の手法を拡張した．Khetan ら [2] は冗長化によって精度を高めた少量の訓練データによって学習器を構築した場合と，冗長化をせずに精度の低い大量の訓練データを与えた場合とで，学習器の性能の変化を調べ，冗長化を行わない方が学習器の性能が高くなることを示した．

本研究は，クラウドソーシングで得たラベルを訓練データとして識別器を構成するという問題設定のもと，様々な条件における最適な冗長度の検証を行う．クラウドワーカの回答集約では，多数決を用いる方法と，Dawid ら [1] の手法を用いた重みつき多数決の 2 手法を用いる．従来の研究では予算の大きさや学習器にとってのタスク難易度が最適な冗長度に与える影響が明らかになっていない．本研究では予算の大きさ，クラウドワーカの回答精度，学習器にとってのタスク難易度を変数として実験を行う．そして実験から導き出される各条件における最適な冗長度を元に，クラウドソーシングによる機械学習器構築における冗長性の意義を検討する．

3　仮説と実験設定

3.1　仮　　説

本研究ではクラウドソーシングにおける冗長化が学習器の予測精度に与える影響について以下に示す 4 つの仮説を立てた．

（1）　予算が大きくなると冗長化のメリットは大きくなる．

（2）　クラウドワーカの回答精度が低くなると冗長化のメリットは大きくなる．

（3）　学習器にとってのタスク難易度が低くなると冗長化のメリットは大きくなる．

（4）　予算、クラウドワーカの回答精度、タスク難易度によって最適な冗長度は変化する．

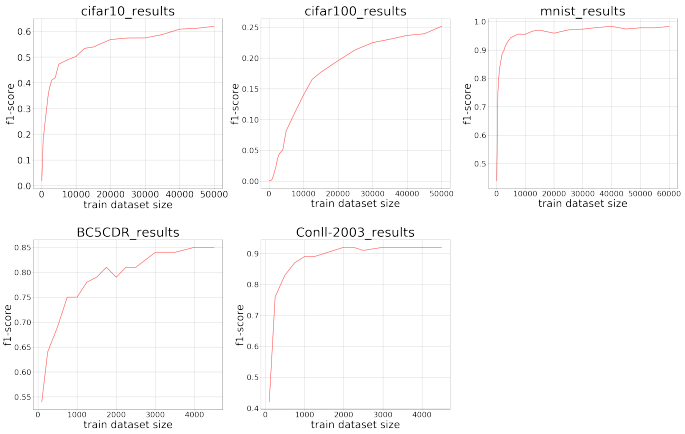

図 1　訓練データ数による学習器の予測精度変化

それぞれの仮説について説明する．１つ目の仮説は，予算が大きくなると冗長化のメリットは大きくなるというものである．訓練データ数が大きくなるにつれ，訓練データ数増加による学習器の予測精度向上量は減少する．このため，予算が大きくなると冗長化のメリットは大きくなると予測する．図 1 は学習器に与える訓練データの量を変化させた時の予測精度の変化を実験によって調べた結果である．用いられているデータの詳細は後述の表 1 に記載している．横軸は学習器のテストデータに対する予測結果の macro-F1 値，横軸は訓練データ数であり，データセットごとの結果をグラフで示す．図 1 より，データセットによって程度の違いはあるものの，ほとんどの場合において訓練データ数が一定以上となったとき急激に予測精度の上昇量が小さくなっていくことが示されている．

２つ目の仮説は，クラウドワーカの回答精度が低くなると冗長化のメリットは大きくなるというものである．クラウドワーカの回答精度が高いときと低いときでは，低いときの方が冗長化によって向上しうる回答精度の上昇幅が大きい．このため，クラウドワーカの回答精度が低くなると冗長化のメリットは大きくなると予測する．

３つ目の仮説は，学習器にとってのタスク難易度が低くなると冗長化のメリットは大きくなるというものである．学習器にとってのタスク難易度が低いとき，少量の訓練データでも十分な訓練が行える．このため，訓練データ数減少によるメリットは大きくなると予測する．

４つ目の仮説は，予算、クラウドワーカの回答精度、タスク難易度によって最適な冗長度は変化するというものである．仮説１～３より，予算，クラウドワーカの回答精度，タスク難易度に応じて冗長化のメリットの大小は変化し，各条件によって最適な冗長度は変化すると予測する．

本研究では，これら 4 つの仮説を実験により検証することでクラウドソーシングを用いた機械学習モデルにおける冗長度の効用を示す．

3.2　実 験 方 法

シミュレーションデータを用いてクラウドソーシングによる学習器構築を再現し，各実験条件において，学習器の性能

（macro-F1 値）を最大化する冗長度の値を明らかにする．クラウドソーシングを再現したシミュレーションデータを訓練データとして学習器に与え，テストデータの予測を行い，予測結果の macro-F1 値を各条件で比較することで学習器の性能を最大化する冗長度の値を明らかにする．F1 値は適合率と再現率の調和平均であり，macro-F1 値は多値分類においてクラスの大きさを考慮しない計算方法によって算出される値である．実験では以下の 3 つの変数を定める．

- 予算の大きさ
- クラウドワーカの回答精度
- 学習器にとってのタスク難易度

予算を冗長度で割った数が，訓練データ数となる．冗長度が 1 の時，予算の大きさが訓練データセットのデータ数となる．実験では，予算の値としてデータセットの訓練データ数の最大値，最大値の $\frac{1}{2}$，最大値の $\frac{1}{10}$ の 3 種類の値を用いて実験を行い，最適な冗長度に与える影響を調査する．

また，実験ではクラウドワーカの回答精度を混同行列によって表現する．ワーカの混同行列の対角成分は，クラウドワーカが正解ラベルと同じラベルを返す確率を表すため，混同行列の値を各実験条件に合わせて定めることでクラウドワーカの回答精度を操作することができる．実験では，クラウドワーカの回答精度として 0.3，0.6，0.9 の 3 種類の値を用いて実験を行い，最適な冗長度に与える影響を調査する．

異なるタスク，異なるデータセットによって実験を行うことでタスク難易度による学習器の振る舞いの変化を調べる．実験では，データセットとして画像分類タスクで 3 つ，固有表現抽出タスクで 2 つの計 5 種類のデータセットを用いて実験を行い，最適な冗長度に与える影響を調査する．

冗長度は同じタスクに対して割り当てるクラウドワーカの人数を示す値である．予算を冗長度で割った値が訓練データセットのデータ数となる．それぞれの実験条件において，冗長度を 1, 5, 20, 50 と変化させて，訓練後の学習器の F1 値に与える影響を調査する．

訓練データ数は学習器に与える訓練データの数を示す．画像分類タスクでは画像とラベル 1 つで 1 つのデータとし，固有表現抽出では 1 文章あるいは数単語列で 1 つのデータとする．

クラウドワーカの回答データ集約には以下の 2 手法による比較を行う．

- 単純多数決
- 重みつき多数決

単純多数決では，各ワーカの票を平等に扱い，多数決を用いてワーカの回答を集約する．重みつき多数決では，ワーカの信頼度によってワーカの回答に重み付けを行った上で多数決を行い，ワーカの回答を集約する．

実験は以下の手順で行う．

（1） データセット，予算の大きさ，冗長度から訓練データ数を決定する．詳細は後述する．

（2） クラウドワーカのシミュレーションデータを作成する．詳細は後述する．

（3） クラウドワーカのシミュレーションデータを多数決あるいは重みつき多数決によって集約し，訓練データセットを作成する．

（4） 訓練データセットを学習器に与えて訓練を行う．

（5） 手順 4 の学習器にテストデータセットの回答を予測させ，予測結果の精度を計測する．

表 1　データセット

名称	タスク種類	ラベル数	訓練データ数
MNIST	画像分類	10	60000
CIFAR10	画像分類	10	50000
CIFAR100	画像分類	100	50000
CoNLL-2003	固有表現抽出	9	4500
BC5CDR	固有表現抽出	5	4500

3.3　実験で扱うタスク

実験では画像分類タスクと固有表現抽出タスクを扱う．使用したデータセットを表 1 に示す．画像分類タスクでは，0-9 の手書き数字の画像データセットである The MNIST database [1]，10 クラスに分類されたオブジェクト認識用の画像データセットである CIFAR10 [10]，100 クラスに分類されたオブジェクト認識用の画像データセットである CIFAR100 [10] の 3 種類のデータセットによる実験を行う．固有表現抽出タスクでは，ニュース記事の一部を切り出した，人名，地名，組織名などを固有表現として定義する CoNLL-2003 [11] と，薬学用語を固有表現として定義する BC5CDR [12] による実験を行う．

タスクごとに機械学習モデルの構築を行う．画像分類では PyTorch のチュートリアル [2] に示された畳み込みネットワークを用いたモデルによって訓練・予測を行う．固有表現抽出では Bi-LTSM CRF [13] を用いたモデルによって訓練・予測を行う．固有表現抽出タスクでは単語ごとにラベルの予測を行う．予測結果については固有表現ごとに正誤の確認をし，固有表現全体が抽出されていたときを正解とする．

3.4　クラウドワーカのシミュレーション

画像分類タスクでは，確率分布から混同行列を作成する．画像分類タスクにおけるシミュレーションワーカの作成は以下のように行う．まずワーカの回答能力を表す混同行列をシミュレーションワーカの人数分作成し，混同行列の値とタスクの真の回答からシミュレーションワーカの回答を生成する．実験条件により定められたクラウドワーカの回答精度 α が平均値となるように，ベータ分布よりサンプルをとったシミュレーションワーカの集合を W とする．C^w をワーカ $w \in W$ の回答精度とする．次に，サンプルされた値が各ワーカの混同行列の対角成分の平均値となるように，各ワーカについて，ベータ分布よりラベル数のサンプルをとる．

ベータ分布 $Sx^{a-1}(1-x)^{b-1}$(S は規格化定数，a，b はパラメータ) はパラメータ b の値を $b = 10$ として固定し，分布の期待値 α が実験条件により定められたクラウドワーカの回答精度

1：http://yann.lecun.com/exdb/mnist

2：https://pytorch.org/tutorials/beginner/blitz/neural_networks_tutorial.html

の平均値となるようにパラメータ a を式 (1) より定めた．

$$a = \frac{\alpha b}{1-\alpha} \tag{1}$$

K をタスクのラベル数とする．各ワーカ $w \in W$ および各ラベル $1 \leqq i \leqq K$ ごとに，上述のベータ分布からサンプル c_i^w を生成し，それぞれをシミュレーションワーカの混同行列の各対角成分の値とする．混同行列の対角成分以外の値は $[0, 1]$ 上の一様分布からサンプリングし，和が $1 - c_i^w$ になるように正規化する．

固有表現抽出タスクでは，実際のクラウドワーカの回答に基づいて混同行列を作成する．固有表現抽出タスクでは，ラベルによって出現頻度が大きく違うという特徴がある．特に，固有表現でないことを示す O タグは正解ラベルの半分以上を占める場合がほとんどである．そのため，O タグの回答精度低下が回答全体に及ぼす影響は極めて大きく，混同行列の他のラベルの値と同等に扱うべきではない．また，固有表現抽出では，固有表現の開始位置を表す B タグと開始位置以外を表す I タグがあり，ワーカの各固有表現における B タグと I タグの回答能力は独立でないと考えるのが自然である．以上の理由から，固有表現抽出において画像分類のように特定の分布からサンプルを得ることでシミュレーションワーカの混同行列を作成することは不適切であると考えた．このため，画像分類タスクではシミュレーションワーカを作成したが，固有表現抽出タスクでは実際のクラウドワーカの回答および正解データからワーカの回答能力を表す混同行列を作成し，回答を生成した．

各タスクに対するワーカの割り当て方について説明する．各ワーカの混同行列の対角成分の平均値をワーカのおおよその回答能力とし，各タスクを回答した各ワーカの回答能力の平均をタスクの期待精度と定義する．全タスクの期待精度の平均値が，実験条件で定めたワーカの回答能力から誤差 0.001 以内となるように，各タスクに対してワーカの割り当てを行った．以上の方法により，データセットからクラウドワーカの回答精度及び冗長度を操作したシミュレーションデータを作成した．

4 実験結果

図 2，図 3 は実験の結果である．図 2 は単純多数決によってワーカの回答集約を行なった結果，図 3 は重み付き多数決によってワーカの回答集約を行なった結果を示す．グラフ群の行はデータセットを示し，列はワーカの回答精度を示す．また，各グラフの横軸は冗長度，縦軸は F1 値を示す．各実験条件について独立に 10 回ずつ実験を行い，実線により結果の平均値，実線と同色の領域により実験データの最大値および最小値を示している．

図 2 と図 3 でデータセット，ワーカの回答精度が同じ実験同士を比較すると，非常に近しい形状のグラフとなった．

予算の変化による影響をみると，ほとんどの実験条件で予算が小さくなるほどグラフの傾きが大きくなっており，予算が小さいときには冗長化のメリットが小さく，予算が大きいときに

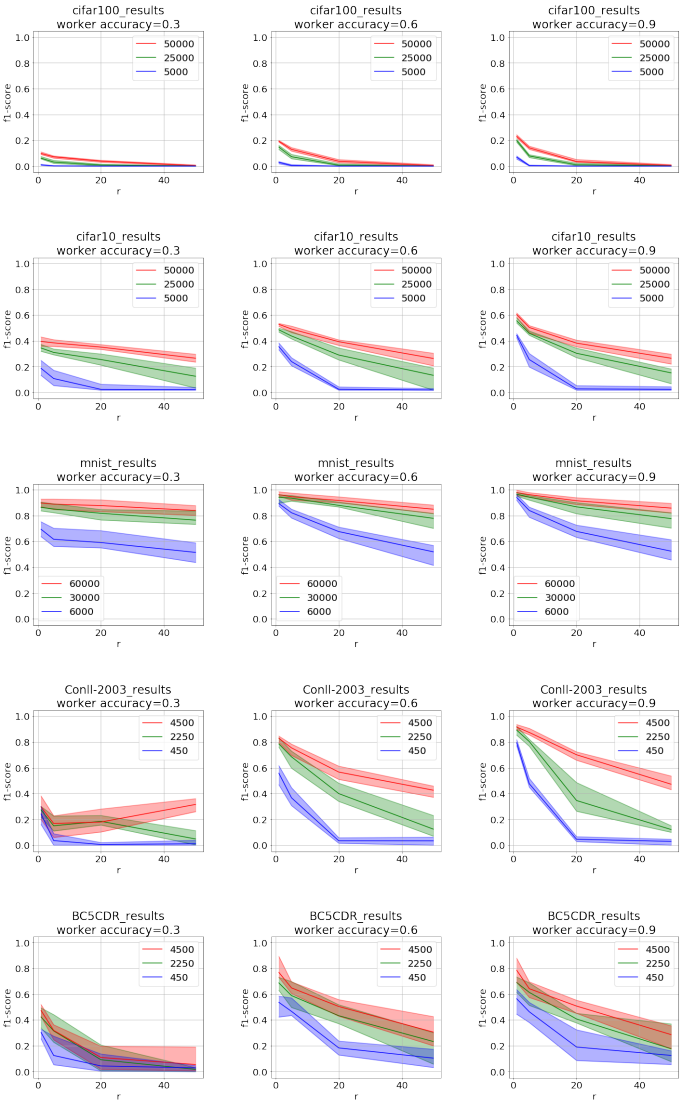

図 2　単純多数決からの訓練結果

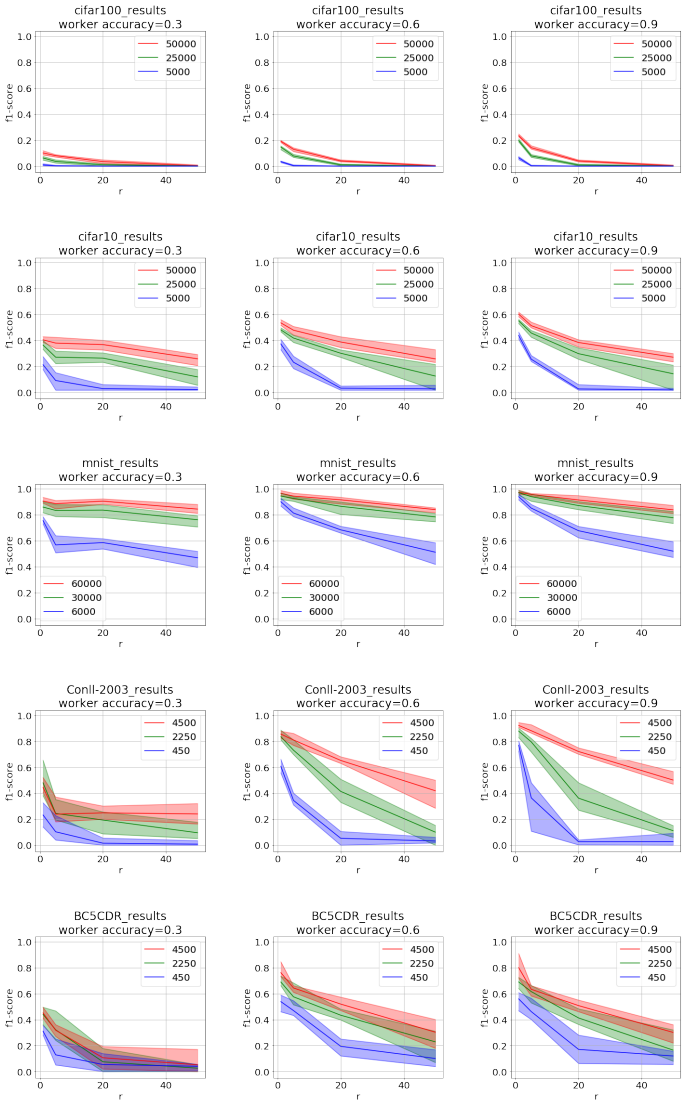

図 3　重みつき多数決からの訓練結果

は相対的に冗長化のメリットが大きくなることが明らかになった．このことから，予算が大きくなると冗長化のメリットは大きくなるとした仮説 1 は正しいことが示された．

クラウドワーカの回答精度の変化による影響をみると，ほとんどの実験条件でクラウドワーカの回答精度が高くなるほどグラフの傾きが大きくなっており，クラウドワーカの回答精度が低いときには冗長化のメリットが大きく，クラウドワーカの回答精度が高いときには相対的に冗長化のメリットが小さくなることが明らかになった．このことから，クラウドワーカの回答精度が低くなると冗長化のメリットは大きくなるとした仮説 2 は正しいことが示された．

データセットごとの特徴を述べる．CIFAR100 を用いた実験では予算の大きさ，ワーカの回答精度に依らず，冗長度が 1 のときが最も優位となった．CIFAR100 を用いた実験ではどの実験条件においても冗長度を上げるほどグラフの傾きが小さくなっている．また，ワーカの回答精度が高くなるほどグラフの傾きが大きく，冗長度を下げるメリットが小さくなっている．

CIFAR10 を用いた実験では予算の大きさ，ワーカの回答精度に依らず，冗長度が 1 のときが最も優位となった．ワーカの回答精度が 0.3，予算の大きさが 50000，25000 の実験では冗長度を上げたときの F1 値の下降量は一定に近い．一方，ワーカの回答精度が 0.9，予算の大きさが 50000，25000 の実験結果では，F1 値の下降量は冗長度を 1 から 5 に変化させたとき最も大きく，冗長度が大きくなるにつれ小さくなっている．このことから，ワーカの回答精度が高くなるほど冗長度を下げるメリットが小さくなっていることが読み取れる．

MNIST を用いた実験では予算の大きさ，ワーカの回答精度に依らず，冗長度が 1 のときが最も優位となった．予算の大きさが 60000，30000 の実験では冗長度を上げたときの F1 値の下降量は一定に近い．予算の大きさが 6000 の実験では，冗長度を上げたときの F1 値の下降量は冗長度を 1 から 5 に変化させたとき最も大きく，冗長度が大きくなるにつれ下降量は小さくなっている．ワーカの回答精度が 0.3，予算の大きさが 60000 のときグラフの傾きは極めて小さい．このため，冗長化のもたらす，訓練データセットの精度向上というメリットとデータ数減少というデメリットが学習器の性能に与える影響が釣り合っている状態に近いことが読み取れる．

CoNLL-2003 を用いた実験では，ワーカの回答精度が 0.3，予算の大きさが 4500 以外の条件では，冗長度が 1 のときが最も優位となった．ワーカの回答精度が 0.9，0.6 の実験では，冗長度を上げたときの F1 値の下降量は冗長度を 1 から 5 に変化させたとき最も大きく，冗長度が大きくなるにつれ下降量は小さくなっている．ワーカの回答精度が 0.3，予算の大きさが 4500 の実験では，冗長度を 1 から 5 に変化させたとき F1 値が下がり，その後冗長度の増加につれて F1 値が大きくなっている．この結果については 5 節で述べる．

BC5CDR を用いた実験では予算の大きさ，ワーカの回答精度に依らず，冗長度が 1 のときが最も優位となった．BC5CDR を用いた実験では，どの実験条件でも冗長度を上げたときの F1 値の下降量は冗長度を 1 から 5 に変化させたとき最も大きく，

冗長度が大きくなるにつれ下降量は小さくなっている．

固有表現抽出タスクである Conll-2003．BC5CDR では結果に大きな違いは見られなかった．画像分類タスクである CIFAR100，CIFAR10，MNIST では，100 種類のカラー画像の分類を行う CIFAR100 が最も学習器にとっての難易度が高く，10 種類の白黒画像の分類を行う MNIST が最も難易度が低いといえる．グラフの傾きも CIFAR100 が最も高く，MNIST が最も小さくなっており，学習器にとってのタスク難易度が高いほどグラフの傾きが大きくなることが明らかになった．このことより，学習器にとってのタスク難易度が低くなると冗長化のメリットは大きくなるとした仮説 3 は正しいことが示された．

図 2，図 3 のほとんどの実験条件において，グラフは右肩下がりになっており，冗長度が 1 の時が最も優位となることが示されている．このことより，予算、クラウドワーカの回答精度、タスク難易度によって最適な冗長度は変化するとした仮説 4 は誤りである．

5 考　　察

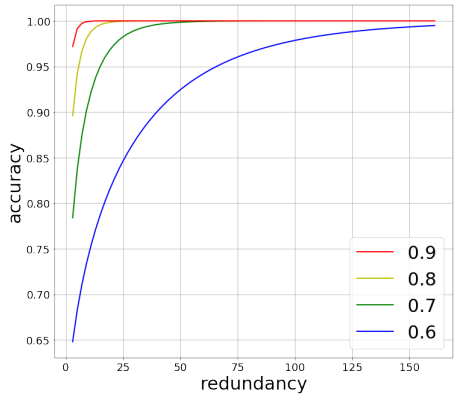

図 4　冗長度による多数決の効用変化

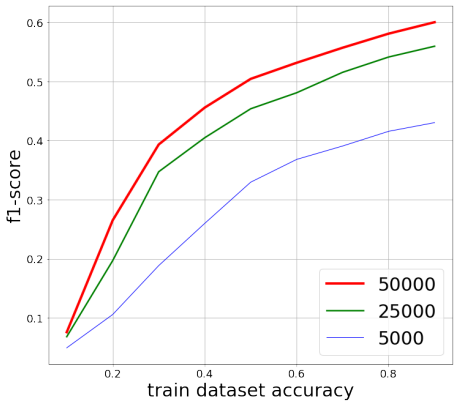

図 5　訓練データ精度による学習器の予測精度変化

図 4，図 5 は実験の結果を考察するための参考データである．図 4 はクラウドワーカの回答精度が一定であると仮定し，二値分類問題に対してワーカの回答で多数決をとった時に正答が得られる確率を算出したものである．縦軸は正答を得られる確率，横軸は冗長度，実線はそれぞれワーカの回答精度ごとの結果を示す．図 5 は学習器に与える訓練データの精度を変化させた時の予測精度の変化を実験によって調べた結果である．横軸は学習器のテストデータに対する予測結果の macro-F1 値，横軸は訓練データの回答精度，実線はそれぞれ訓練データごとの結果を示す．

図 2，図 3 より，予算の大きさが小さいほど冗長度を上げるメリットが小さくなっていることが示された．これは訓練データ量が少ないほど訓練データ量増加の恩恵が大きく，予算の大きさが小さいほど冗長度を下げることによるデータ量の相対的増加のメリットが顕著に現れたためと考えられる．

また，図 2，図 3 においてワーカの回答精度が高くなるほど冗長度を上げるメリットが小さくなっていることが示されている．

ワーカの回答精度が高い状況では少ない人数で正解率が収束し，収束した後は冗長度増加の恩恵がほとんど見られないことが図 4 より示されている．さらに，図 5 より，回答精度が高くなるほど回答精度増加の恩恵が小さくなることが示されている．このことから，ワーカの回答精度が高くなるほど冗長度を上げるメリットが小さくなったと考えられる．

CoNLL-2003 を用いた実験では．予算の大きさが 4500，ワーカの回答精度が 0.3 のとき，冗長度を 1 から 5 に変化させたとき急激に F1 値が下がり，その後冗長度の増加につれて F1 値が大きくなる結果となった．

CoNLL-2003 は，図 1 より，訓練データ数が一定の値になるまでは急速に予測精度が上昇し，その後は予測精度の伸びが極めて小さくなることが示されている．このため，CoNLL-2003 を用いた潤沢な予算設定での実験では，冗長度を上げることによるメリットが大きい．

また，図 5 より，訓練データの精度が低いときほど精度上昇による F1 値の上昇量が大きくなることが示されている．固有表現抽出はフレーズ全体をラベル付する必要があるため，固有表現の前後のラベルに誤りがある場合や固有表現の一部しかラベル付されていない場合には，その固有表現全体が誤りと判定される．今回の実験ではシミュレーションワーカの回答は単語ごとに作成されているため，作成された回答データの精度は設定したワーカの回答精度より低くなる．このため，CoNLL-2003 でワーカの回答精度を 0.3 に設定した実験では回答データの精度が極めて低く，訓練データの精度上昇によるメリットが大きくなった．

これらの理由から，CoNLL-2003 を用いた予算の大きさが 4500，ワーカの回答精度が 0.3 の実験では，冗長度を増加につれて F1 値が大きくなる結果となった．

さらに，ワーカの回答精度が低いとき，低い冗長度では冗長度増加による恩恵があまり得られないため，CoNLL-2003 を用いた予算の大きさが 4500，ワーカの回答精度が 0.3 の実験では，冗長度を 1 から 5 に変化させたとき急激に F1 値が減少する結果となったと考えられる．

機械学習は訓練データ全体の傾向を訓練することで，予測性能を獲得する．よって，学習器の訓練自体がワーカの回答全体に対して擬似的に多数決をとっているものとして解釈すること

ができる．このことから，多数決のような統計的手段で解決できるモデルが確立されているタスクについては，あらかじめ多数決を行なったデータを与える必要は無いと考えられる．

6 おわりに

本論文では，クラウドワーカの回答精度，学習器にとってのタスク難易度に関わらず，冗長度が 1 の時最も学習器の予測精度が高くなることが示された．また，予算の大きさが小さい，ワーカの回答精度が高い，学習器にとってのタスク難易度が高いという条件のとき，特に冗長度を下げる恩恵が大きいことが実験より明らかとなった．

今回の検証では画像分類タスクと固有表現抽出タスクについて扱ったが，クラウドソーシングによる訓練データ作成は翻訳タスクや要約タスクでも使われている．これらのタスクにおいて冗長化がどのような意義を持つかの検証を今後は検討していきたい．

謝　辞

本研究の一部は，JSPS 科研費（課題番号 19K20333）および JST CREST (JPMJCR16E3) AIP チャレンジの助成によって行われた．

文　献

[1] P. Dawid, A. M. Skene, A. P. Dawidt, and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, pp. 20–28, 1979.

[2] Ashish Khetan, Zachary C. Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. In *International Conference on Learning Representations*, 2018.

[3] Victor S. Sheng, Foster J. Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In Ying Li, Bing Liu, and Sunita Sarawagi, editors, *KDD*, pp. 614–622. ACM, 2008.

[4] Edwin Simpson and Iryna Gurevych. Bayesian ensembles of crowds and deep learners for sequence tagging. *CoRR*, Vol. abs/1811.00780, , 2018.

[5] An Thanh Nguyen, Byron Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 299–309, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[6] Matthew Lease. On quality control and machine learning in crowdsourcing. In *Human Computation, Papers from the 2011 AAAI Workshop, San Francisco, California, USA, August 8, 2011*, Vol. WS-11-11 of *AAAI Workshops*. AAAI, 2011.

[7] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[8] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition, 2016.

[9] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[10] Krizhevsky, A., Hinton, G. Learning multiple layersof features from tiny images(Technical Report). Master's thesis, Department of Computer Science, University of Toronto, 2009.

[11] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003.

[12] Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. Learning named entity tagger using domain-specific dictionary. In Proc. EMNLP, pp. 2054–2064, 2018.

[13] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition, 2016.
beginner/blitz/neural_networks_tutorial.htmlsphx-glr-beginner-blitz-neural-networks-tutorial-py, (accessed 2021-2-11).