主観を要するアノテーションタスクの観察と可視化

三浦 梨花† 栃木 彩実† 伊藤 貴之†

†お茶の水女子大学理学部 〒112-8640 東京都文京区大塚 2-1-1

E-mail: †{g1820536,itot}@is.ocha.ac.jp

あらまし アノテーションは機械学習の上流プロセスであり，このプロセスを経た訓練データが学習に用いられる．よってアノテーションの信頼性は極めて重要である．ここでアノテーションには作業者（ワーカ）ごとの傾向があり，これらの傾向差がデータの信頼性を損ねる可能性がある．特にワーカの主観に依存するタスクにおいて，ワーカごとの傾向差は顕著に現れる．そこで本研究ではワーカのアノテーション結果を観察することで，信頼性の高いアノテーションの実現を目指す．具体例として本報告では，977 枚の顔画像に 3 人のワーカがリッカード尺度を用いて表情を評価したアノテーションを題材として，可視化結果からその信頼性を検証する．

キーワード 主観，可視化，アノテーション，品質管理，訓練データ

1. はじめに

機械学習の精度や信頼性を向上するには，上流プロセスであるアノテーション作業の信頼性を担保することで訓練データの品質を向上させることが重要である．一方で，訓練データの作成は人手で行われることが多い．そのため，ワーカの能力や知識，育った環境などの属人的な要因からデータにばらつきやブレが生じてしまう可能性がある[1]．データのばらつきやブレは，訓練データの品質に大きな影響を与える．特に，年齢推定や性別推定などに代表される，ワーカの主観によってアノテーションされるタスクの際は，明確な判断基準がないことや，各個人のアノテーションの判断も一貫しているとは限らないことなどの理由[1]により，ばらつきやブレが大きくなり，信頼性の高い訓練データを得ることが難しい場合がある．

このようなデータのばらつきやブレにはワーカごとの傾向がある．例として，人物画像に対して 8 人の作業者が年齢推定結果を注釈した結果を可視化した村上らの研究[2]では，40～60 代の人物画像に対して高めの年齢推定をする作業者や，逆に低めの年齢推定をする作業者がいることがわかった．また，人物の感情を推定するタスクの際にも，ワーカの感情経験などの要因により，感情のラベルや極性が個人によって異なることが示されている[3][4]．

データのばらつきやぶれ，正解率などからアノテーションの質についての議論は多くされてきたが，ワーカのアノテーション傾向についてはあまり議論されてこなかった．しかし，訓練データの品質を向上させるためには，訓練データを作成しているワーカの傾向を検証することも重要な課題である．アノテーションの工程が人手によることから，作業の疲れによってアノテーションの判断基準にブレが生じることや，作業に慣れることでワーカ個人の判断基準が明確になるなど，時間的な要因による訓練データの信頼性の変化も想定される．このように，ワーカの作業時間とデータの品質関係を検証することで，アノテーション作業の時間的要因からおこる問題の改善点を見つけ出すことができ，より効率的に信頼性の高い訓練データを得ることができると考える．また，ワーカごとのアノテーション傾向や信頼性を可視化し理解しやすくすることで，アノテーション結果に及ぼした影響を考慮することができ，データの作成に有効活用することができる可能性がある．ここから本報告では，訓練データをワーカの観点から分析・説明することで，信頼性の高いアノテーションを実現することを目指す．

本報告では個人の主観に依存するタスクとして，顔画像 977 枚にリッカード尺度を用いて 6 つの表情の項目に対してアノテーションするタスクを 3 人のワーカに依頼し分析した結果を紹介する．本研究の主な貢献は以下の通りである．

- アノテーションの所要時間・経過時間・データの品質の関係を観察したこと．
- ワーカごとのアノテーション傾向を可視化することで，アノテーションの難しい項目を見つけることができる可能性を示唆したこと．
- アノテーションの誤りにもワーカごとに傾向が存在する可能性があることを発見したこと．

本報告の構成は以下の通りである．2 章では関連研究について，3 章では提案手法について，4 章では実行結果について述べる．そして，5 章では分析結果から得た考察について，6 章では本報告のまとめと今後の課題について述べる．

2. 関連研究

本章では関連研究として，訓練データの信頼性の評価手法と訓練データの可視化手法についての論文を紹介する．また，関連研究を示すことで，本研究の差異性と位置付けを述べる．

2.1 訓練データの信頼性評価

アノテーションの信頼性を評価することで，訓練データの信頼性の向上を目指す研究はいくつかある．Dawid ら[5]は，ワーカの誤差を測定する問題について EM アルゴリズムを用いて真の回答の推定とワーカの能力の推定を交互に行うことで信頼度の高いデータを得るモデルを提案している．また光田ら[6]は，ワーカの作業中の視線情報を信頼性の推定に用いることで，より正確にアノテーションの信頼性を分析し，訓練データの信頼性の向上に取り組んだ．駒谷ら[7]は，アノテーションの信頼性を評価する際に Fleiss's κ [8] ，Krippendorff's α [9]の 2 つの指標を導入して比較した．その結果, Krippendorff's α を用いたほうが，不一致の度合いが考慮されてワーカ間の一致率が極端に低くならないことを示した．本研究でもこれと類似したアプローチで信頼性を評価している．対して，これらの研究では，アノテーション作業時間とデータの信頼性の関係の分析や，ワーカ個人のアノテーション項目の信頼性の関係の分析に取り組んでいない．そこで本研究では，ワーカのアノテーション作業中に所要時間や経過時間のログをとることで，時間とデータの信頼性との関係について議論する．また，項目ごとにワーカの信頼性を評価することで，アノテーションの傾向を分析する．

2.2 アノテーション傾向の可視化手法

村上ら[2]は，ワーカ間のアノテーションの不一致をヒートマップで可視化するツールを開発した．その結果として，訓練データが主観的な要素に依存するタスクにおいては，ワーカごとにアノテーションの傾向があることを示した．駒谷ら[7]は，アノテーション結果の傾向を理解するために，ワーカ間での訓練データのスコアの付与傾向を混同行列と回帰分析により分析した．稲垣ら[10]は，ワーカが回答したデータを多次元尺度構成法によって可視化し、スペクトラルクラスタリングを用いて分類することで，少数の特徴的な回答をするワーカを抽出した．

以上の研究では訓練データの全体的な傾向の可視化にとどまっており，個々のワーカの項目ごとの可視化，データの信頼性と可視化結果の関係，といった点まで含めて包括的に取り組んだ研究はほとんどない．そこで本研究では，ワーカごとに訓練データを可視化し，信頼性評価値と合わせて分析することで，より深く各ワーカのアノテーション傾向について分析する．

3. 提案手法

本章では提案手法の処理手順を述べる．3.1 節では本研究で使用するデータについて説明し，続いて主観を要するタスクのアノテーション実施手順について説明する．3.2 節では信頼性の算出手法について，3.3 節ではアノテーション作業経過枚数と所要時間の関係の分析手法について，そして 3.4 節では多次元の訓練データの可視化手法について説明する．

3.1 使用するデータと主観評価の収集

本研究ではデータセットとして，顔表情データベース FACES database を使用した[11] ．このデータセットは 58 人の若年者，56 人の中年者，57 人の高齢者の計 171 名が参加して，それぞれ 6 つの表情(happiness, disgust，anger，neutrality，sadness and fear)を提供しているデータベースである．このデータベースに含まれる 977 枚の顔画像を対象として，20 代の女性 3 名に，5 段階のリッカード尺度で主観評価を依頼し実施した．印象評価の項目は表 1 の通りである．

表 1 評価項目

	全くそう思わない			とてもそう思う
	1	2	3	4	5
happiness	├	┼	┼	┼	┤
disgust	├	┼	┼	┼	┤
anger	├	┼	┼	┼	┤
neutrality	├	┼	┼	┼	┤
sadness	├	┼	┼	┼	┤
fear	├	┼	┼	┼	┤

3.2 信頼性の評価指標

本研究では，主観にもとづいてアノテーションされたタスクを扱うため，複数のワーカによるアノテーション結果が一致するとは限らない[7]．そこで，ワーカ間の一致度からアノテーションの信頼性を分析する．アノテーションの信頼性を評価する手法として，本研究では Krippendorff's α を用いる．また本研究では，算出した α 値の妥当性について，級内相関係数(ICC)を適用して信頼性評価値を算出することで検証する．両者ともに評価者間信頼性を測る尺度である．評定者間信頼性(inter-rater reliability)の代表的な指標として他にも Kendall の一致係数やカッパ係数[8]などの種類があるが，本研究で用いる訓練データはリッカード尺度を用いているため，尺度水準を間隔尺度とみなせることを考慮して計算することのできるこの 2 つの指標を採用した．以下，この 2 つの指標について説明する．

Krippendorff's α :

2 人以上のワーカ間の一致度を計算する指標．スコア間距離を尺度水準によって変更することができるた

め, 高い汎用性を持つとされている[12] . 本研究では, リッカード尺度でアノテーションされた画像群を使用することから, R の irr パッケージを用いて Interval 指標で Krippendorff's α を計算する.

先行研究[7]で言及されているように, 社会学の研究では一般にα>0.8 が信頼性を保った一致率であるとされているが, 各ワーカの主観によってアノテーションされるタスクではアノテーションの一致率は低くなる傾向にあり, このようなタスクではαの値は 0.4 程度が妥当であるとされている.

級内相関係数(ICC):

級内相関係数は, 検者内または検者間の信頼性を求める手法である[14] . 級内相関係数には Case1, Case2, Case3 の 3 種類があるが, 本研究では特定の検者の検者間信頼性を求めたいため, R の psych パッケージの関数 ICC()で Case3 を利用して計算する. ICC の判定基準として以下の表 2 を採用する. しかしこの表は, Landis ら[13]による Kappa 係数の表を ICC の判定に応用したものであり, 理論的根拠はないと対馬は述べている[14] .

表 2 級内相関係数判定基準

級内相関係数	判定
0.0~0.2	Slight
0.21~0.40	Fair
0.41~0.60	Moderate
0.61~0.80	Substantial
0.81~1.00	almost perfect

3.3 時間とデータの品質の関係性分析手法

アノテーション作業の経過枚数と顔画像 1 枚に対するアノテーション所要時間が, どのように訓練データの品質に関係してくるのかを把握するため, 可視化を適用してアノテーション結果を観察する. 処理手順は以下の通りである.

① 経過枚数と顔画像一枚あたりのアノテーション所要時間はデータの単位が違うので, 尺度を揃えるためにデータの前処理として Python の scikit-learn ライブラリの StandardScaler によって標準化を施す.

② 標準化を施したデータに対して, Python のモジュール Scipy で階層型クラスタリングを行う. 本研究では, クラスタの生成方法としてウォード法, 基準となる距離としてユークリッド距離を用いる.

③ 考察するクラスタ数の範囲を 2~25 として, 表 3 に示した 26 個の指標を用いることでクラスタリングの最適クラスタ数を決定する. 指標と最適クラスタ数の計算に本研究では Python の Nbclust package[15]を用いている. 判定された結果からクラスタ数を 17 とした.

表 3 最適クラスタ数の決定に用いる 26 指標

ch	duda	pseudot2	cindex	beale
ccc	ptbiserial	db	frey	hartigan
ratkowsky	scott	marriot	ball	trcovw
tracew	friedman	mcclain	rubin	kl
silhouette	dindex	dunn	hubert	sdindex
sdbw				

④ クラスタごとに色を割り当て散布図で可視化する.

⑤ 3.2 節で述べた 2 つの指標を用いてクラスタごとに信頼性評価値を算出し, 可視化結果と信頼性評価値を照合する. これにより, 経過時間・所要時間・データの信頼性の関係を考察する.

3.4 ワーカのアノテーション傾向可視化手法

本節では, 多次元データである訓練データを可視化する 2 種類の手法について述べる.

3.4.1 PCA(主成分分析)での可視化

各ワーカのアノテーション傾向を把握するために, 多次元の訓練データを PCA で可視化する. この可視化によって, 各項目にどのような傾向があるのかを短時間で観察できる.

3.4.2 平行座標プロットでの可視化

PCA でおおまかに観察した各項目にどのような傾向があるのかを具体的に考察するために, 平行座標プロットで可視化する. これによって特定の画像・特定のワーカについて細かく観察し, より深く考察することができる.

4. 実行結果・考察

本章では実行結果から得られる知見について述べる. 4.1 節ではアノテーションの時間と訓練データの品質の関係について, 4.2 節ではワーカごとのアノテーション傾向について論じる.

4.1 分析結果①-時間変化の関係-

3.3 節に示した手順に沿って, Rstudio の shiny パッケージで可視化した結果を示す. 作成した散布図を図 1 に示す. 点 1 個が画像 1 枚に相当しており, 多次元データは 6 個の項目を 6 次元データにしたものである. 3.3 節の手法において決定した 17 個のクラスタの各々に固有の色を割り当てている. また, 横軸はアノテーションした顔画像の経過枚数, 縦軸は一枚の顔画像をアノテーションするのに要した時間である. クラスタごとのα値と級内相関係を計算した結果を, それぞれ表 5, 6 に示す. 各項目の相関係数は, 表 4 の通りとなった. α値と級内相関係数には強い正の相関があることから, α値には十分な信頼性があるとみなし, 図 1 の散布図と表 1 のα値を用いた. なお, 17 個のクラス

タのうち，クラスタ内の顔画像の枚数が 3 枚以下であった cluster5, cluster6, cluster7 は除外して出力している．

表 4　α 値と ICC の相関係数

happiness	0.99978494
disgust	0.94844305
anger	0.91282669
neutrality	0.99526125
sadness	0.99586303
fear	0.98475636

表 5 において，α 値が低く出ている cluster10, cluster11 を図 2 に示す．この 2 つのクラスタはアノテーション作業の序盤でかつ，アノテーションに時間がかかっている画像群である．対して，α 値が高く出ている cluster1, cluster2, cluster4 を図 3 に示す．これらのクラスタはアノテーション作業の終盤でかつ，アノテーションの所要時間が短い画像群である．図 2 と図 3 を比較した結果から，データの信頼性はアノテーション開始時よりも終盤の方が高く，そして所要時間は長いよりも短い方が高くなる傾向があることがわかる．この結果から，作業の序盤では，アノテーションの判断基準が定まっていないためにブレが生じること．また，アノテーションに時間がかかるタスクは，アノテーションの判断が難しいタスクであり信頼性も低くなることが推測される．

アノテーション作業終了後にワーカに対して実施したアンケートにおいて,「アノテーション序盤・中盤・終盤を比べた時にアノテーションの正確さはどれが高いと感じるか」という質問に対して，3 人中 2 人がアノテーション終盤であると答えている．その理由として，「自分の中で判断基準が明確になってきたから.」「怒りと嫌悪の違いについて最初はわからなかったが，鼻をしわくちゃにさせるのが嫌悪だと気付いてからはその 2 つの違いがわかるようになった.」と回答している．ここからも，作業終盤はアノテーションに慣れていることで，判断基準がワーカの中で確かになり，ブレが生じにくくなっていると考察できる．一方で，開始時の方がアノテーションの正確性が高いと回答した 1 人は,「最初の方が集中してじっくり考えたから.」と述べている．

また，「作業の疲労感によりアノテーションの精度は下がったと思いますか.」という質問に対して 3 人中 2 人のワーカが「はい」と答えている．このように，アノテーション作業の疲労感からくるアノテーションの精度の低下が理由にあげられていたことから，作業の疲れによってデータの品質に影響が出てくることが示唆される．よって，ワーカによっては休憩を取るなどして，アノテーション環境を整えることで，訓練データの信頼性を向上できる可能性があると考えられる．

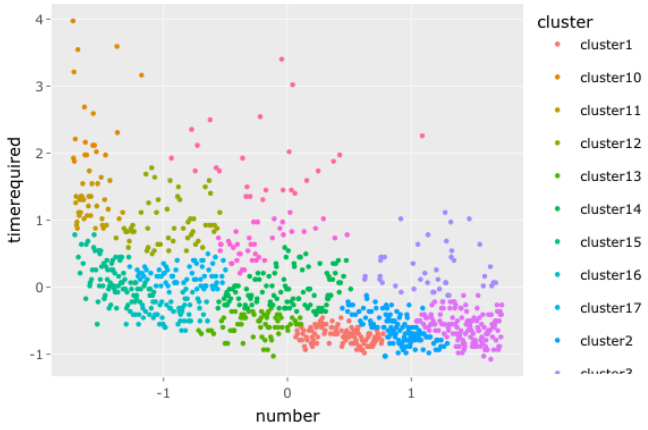

図 1　経過枚数と所要時間を表す散布図

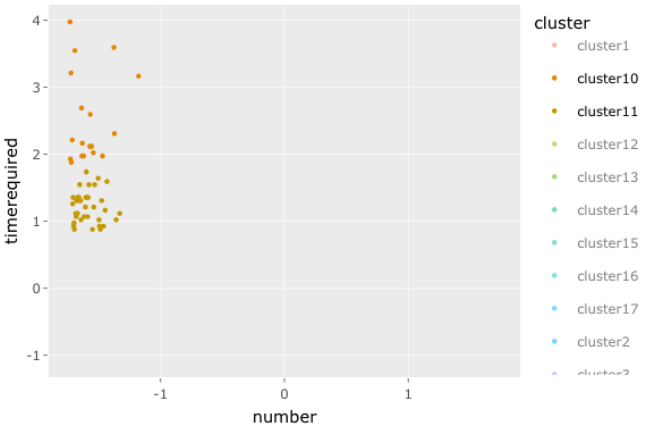

図 2　α 値が低く出ているクラスタ

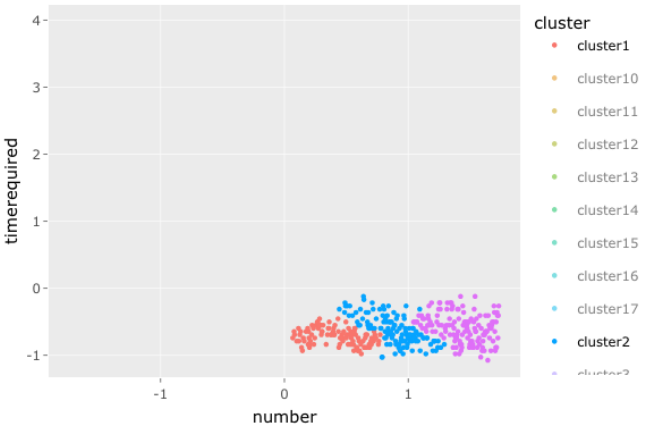

図 3　α 値が高く出ているクラスタ

表 5 クラスタごとの α 値

α	happiness	disgust	anger	neutrality	Sadness	fear
cluster1	0.967724	0.485361	0.403226	0.793166	0.817601	0.728619
cluster2	0.965039	0.551477	0.22094	0.786552	0.768749	0.589559
cluster3	0.963677	0.550353	0.44836	0.651242	0.610709	0.591247
cluster4	0.913324	0.618843	0.302452	0.760678	0.71124	0.637917
cluster8	0.902836	0.337491	0.028397	0.723826	0.638271	0.340086
cluster9	0.929871	0.275892	0.028397	0.781872	0.525529	0.521925
cluster10	0.12135	0.255084	0.106886	0.065308	-0.000395	0.130081
cluster11	0.538456	0.25341	0.534819	0.142251	0.282937	0.056511
cluster12	0.692826	0.443844	0.321448	0.520011	0.358001	0.253021
cluster13	0.963477	0.494209	0.051478	0.847736	0.784295	0.438707
cluster14	0.97606	0.509429	0.340948	0.794901	0.688034	0.524391
cluster15	0.920293	0.426179	0.396826	0.368471	0.103843	0.429044
cluster16	0.946383	0.417885	0.426925	0.685733	0.501363	0.48539
cluster17	0.79928	0.538314	0.317611	0.723902	0.442791	0.488503

表 6 クラスタごとの級内相関係数

icc	happiness	disgust	anger	neutrality	sadness	fear
cluster1	0.97	0.55	0.52	0.8	0.82	0.76
cluster2	0.97	0.59	0.35	0.8	0.77	0.63
cluster3	0.96	0.56	0.51	0.71	0.62	0.63
cluster4	0.91	0.62	0.39	0.77	0.72	0.66
cluster8	0.91	0.49	0.19	0.74	0.64	0.43
cluster9	0.94	0.35	0.32	0.79	0.54	0.6
cluster10	0.14	0.33	0.24	0.088	0.1636	0.33
cluster11	0.55	0.33	0.55	0.23	0.33	0.191
cluster12	0.71	0.49	0.37	0.52	0.42	0.34
cluster13	0.96	0.56	0.215	0.85	0.79	0.47
cluster14	0.98	0.6	0.5	0.8	0.69	0.59
cluster15	0.92	0.48	0.43	0.45	0.2	0.47
cluster16	0.95	0.45	0.43	0.7	0.53	0.53
cluster17	0.81	0.54	0.36	0.73	0.46	0.56

4.2 分析結果②-ワーカのアノテーション傾向の可視化-

本節では，各ワーカのアノテーション傾向を，主成分分析(PCA)と平行座標プロットの 2 種類の手法で可視化した結果を示す．

4.2.1 主成分分析(PCA)での可視化

Rstudio の shiny パッケージにおいて PCA を利用して各ワーカの多次元の訓練データを可視化した結果を図 4～6 に示す．各表情の画像群ごとに点の色を変えて出力している．図 4～6 において矢印は因子負荷量を表しており，1 つの点は各画像の主成分得点をプロットしたものである．また，ワーカごとに 6 つの項目ごとに α 値を算出した結果を表 7 に示す．これらの結果から読み取れる各ワーカのアノテーション傾向を以下に述べる．

まず，図 4 において因子負荷量の矢印線の向きと長さが同じことから，ワーカ A において項目 fear と項目 sadness は，ほぼ同じ意味をもつと言える．ここからワーカ A は顔画像一枚に対して fear と sadness のアノテーションに関してブレが生じていると推測することができる．そして，fear と sadness，あるいは disgust と anger のように，同様にアノテーションする傾向がある項目については，表 7 からもわかるように信頼性が低くなる傾向にある．また，ここで表 7 を用いてワーカ A の fear と sadness の α 値と比べて観察してみると，この 2 項目の α 値の差が小さいことがわかる．図 5, 6 で観察した場合でも，ワーカ B の anger と disgust などで同様の傾向が観察できた．さらに，因子負荷量が近い 2 つの項目は α の値も近くなる可能性があることがわかった．

続いて，3 人のワーカの共通点について観察した．3 人とも disgust と anger の因子負荷量が近くなっており，かつ，disgust と anger の顔画像が同じような位置に分布している．このことから，disgust と anger の 2 項目は，どのワーカでも混同しやすい傾向があり，3 人

とも disgust と anger のアノテーションにブレが生じやすいことが示唆される．実際に表 7 において disgust と anger の α 値を観察してみると，2 つの変数はどちらも α 値が低く出ており，信頼性が低くなっている．この結果から disgust と anger を適切に区別してアノテーションするのは難しいと推測される．それに対して，happiness, neutrality の 2 変数に関しては，3 人とも他の項目との区別ができており，分布からも他の項目に比べて明確に区別していることを観察できる．以上により，happiness と neutrality の 2 項目はアノテーションが比較的容易であると推測できる．この 2 変数は α 値が比較的高いことが表 7 からもわかる．

表 7 ワーカごと項目ごとに算出した α 値

	A	B	C
happiness	0.938022	0.9330505	0.9405905
disgust	0.2804685	0.404644	0.2539515
anger	0.31344	0.3980485	0.3469255
neutrality	0.6726515	0.698883	0.7368405
sadness	0.537156	0.61828	0.593601
fear	0.5141065	0.5363715	0.469491

また，アノテーション作業後に実施したアンケートの中で，3 人のワーカに，判断しやすい表情を順位づけしてもらった．その結果を表 8 に示す．3 人とも最も判断しやすいと評価したのは happiness の項目であった．また，3 人中 2 人が neutrality を 2 番目に判断しやすいと回答した．表 7 からわかるように，happiness と neutrality は α 値からみても判断しやすいといえる項目であり，ワーカが難しいと感じた項目は信頼性も下がると分析できる．ここからワーカの所感からもデータの信頼性を推察できる可能性があると言える．

表 8 判断しやすかった表情を順位付けした結果

	A	B	C
1 番	happiness	happiness	happiness
2 番	neutrality	neutrality	sadness
3 番	sadness	sadness	disgust
4 番	anger	fear	anger
5 番	fear	disgust	fear
6 番	disgust	anger	neutrality

ここまでに示したように，各ワーカのアノテーション傾向を観察することや，その観察結果と α 値との関係を照合することで，アノテーションが難しい項目を推測することができる．そして，アノテーションが難しいと判断される項目に対して優先的にアノテーションのやり直しを実施することで，訓練データの信頼性の向上を試みることができる．特定のワーカのみが苦手とする項目や得意とする項目に関しては，ワーカ間の評価の平均化の際に重みをつけるなどにより，訓練データの信頼性を向上できる可能性がある．

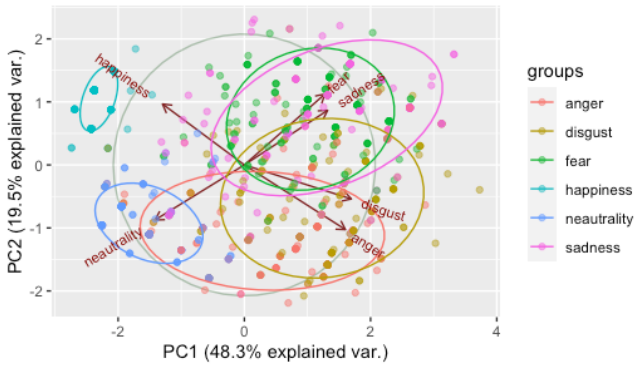

図 4　PCA でのワーカ A の訓練データ可視化結果

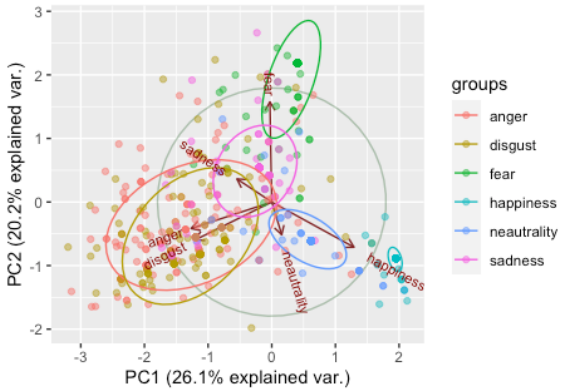

図 5　PCA でのワーカ B の訓練データ可視化結果

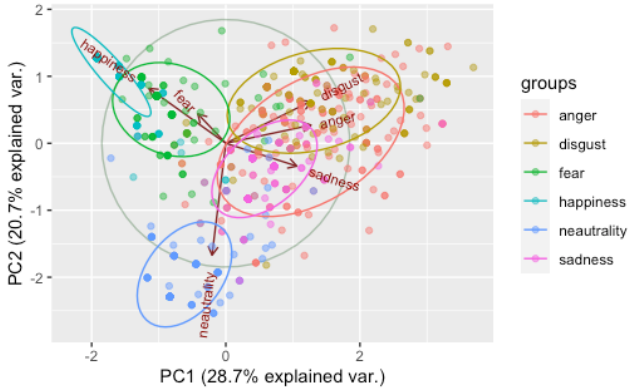

図 6　PCA でのワーカ C の訓練データ可視化結果

4.2.2 Hiplot での可視化

平行座標プロットで訓練データを可視化することで，特定の画像の傾向やばらつきについて観察する．本研究では，Facebook 社が公開している HiPlot[16]というライブラリで訓練データを可視化した．本報告では，可視化結果から観察できた特徴的な傾向の例を紹介する．図 7，図 9 において，折れ線が画像に相当してい

て，7 本の垂直な軸が 6 つの項目と表情ごとのクラスタに対応している．軸をドラッグすることで，指定した範囲の画像のアノテーション結果を出力する．
また，図 8，図 10 はそれぞれ図 7，図 9 の平行座標プロットにおいて選択した範囲の画像群のアノテーション結果をデータテーブルとして出力したものである．

図 7〜10 はそれぞれ，ワーカ A とワーカ C が anger に属する顔画像にアノテーションする際に，anger の項目を「1:全くそう思わない」または「2:そう思わない」と評価した画像群を選択して，それらの画像の各項目への評価を可視化したものである．図 7, 8 から，ワーカ A は anger に属するにもかかわらず「そう思わない」とアノテーションした画像に対して，neutrality であると評価する傾向があることがわかる．一方で図 9, 10 から，ワーカ C は anger に属するにもかかわらず「そう思わない」とアノテーションした画像に対して，sadness であると評価する傾向があることがわかる．このように，本来の想定とは異なるタグがつけられる画像についても，ワーカごとの傾向があることがわかる．

このように，特定の条件を満たす画像に絞って評価結果を可視化し，詳細に観察することで，ワーカごと，あるいは項目ごとの傾向を把握することができる．

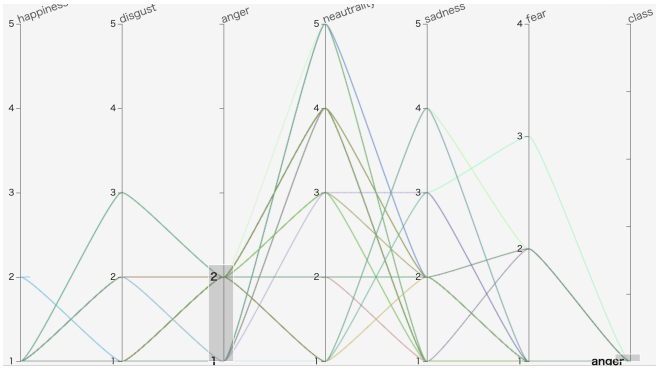

図 7 Hiplot によるワーカ A のデータ可視化結果

	uid	from_uid	happiness	disgust	anger	neautrality	sadness	fear	class
	0	null	2	1	1	5	2	1	anger
	101	null	1	3	2	3	1	1	anger
	102	null	1	2	2	1	2	1	anger
	107	null	1	1	1	5	1	2	anger
	109	null	1	2	2	1	3	3	anger
	113	null	1	1	2	4	1	1	anger
	117	null	2	1	1	4	1	1	anger
	119	null	1	1	2	4	1	1	anger
	120	null	1	1	1	3	3	1	anger
	124	null	1	1	2	4	1	1	anger
	125	null	1	1	1	5	1	1	anger
	130	null	1	1	1	5	1	1	anger
	131	null	1	1	2	4	1	1	anger
	133	null	1	2	2	4	2	1	anger
	136	null	1	1	2	3	1	1	anger
	143	null	1	2	2	1	3	1	anger
	145	null	1	1	2	3	2	1	anger
	15	null	1	2	1	1	4	1	anger
	155	null	1	1	2	4	1	1	anger

図 8 図 7 で選択した項目のデータテーブル

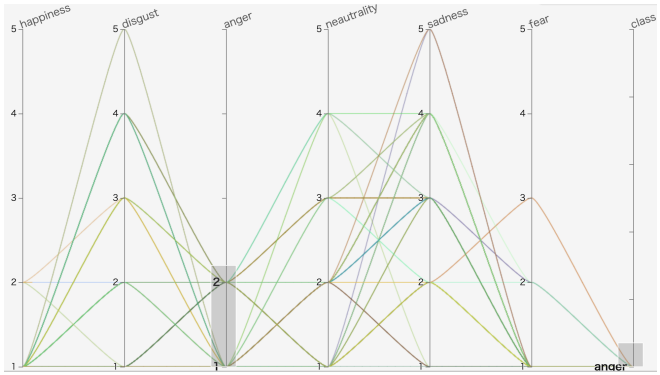

図 9 Hiplot によるワーカ C のデータ可視化結果

	uid	from_uid	happiness	disgust	anger	neautrality	sadness	fear	class
	0	null	2	2	2	2	3	2	anger
	101	null	1	4	2	2	4	1	anger
	102	null	1	5	1	2	3	1	anger
	107	null	1	4	1	2	3	1	anger
	108	null	1	1	2	4	3	1	anger
	109	null	1	4	1	2	3	1	anger
	11	null	1	5	1	1	2	3	anger
	111	null	1	2	2	2	2	1	anger
	113	null	1	4	2	2	4	1	anger
	115	null	1	2	2	2	4	1	anger
	117	null	1	1	1	1	4	1	anger
	124	null	1	1	1	2	4	1	anger
	125	null	1	1	1	2	4	1	anger
	130	null	1	1	1	2	2	3	anger
	132	null	1	1	1	3	2	2	anger
	133	null	1	2	2	3	4	1	anger
	136	null	1	1	1	3	3	1	anger
	143	null	1	1	1	1	5	1	anger
	152	null	1	1	1	1	4	1	anger

図 10 図 9 で選択した項目のデータテーブル

5. 結論・今後の課題

本報告では，主観を要するタスクによって構築された訓練データを可視化する手法と，そのデータの信頼性との関係をワーカの観点から考察する方法を提案し，信頼性の高いアノテーションを実現する手段について議論した．その結果として以下のことがわかった．

- 訓練データのアノテーションのブレやばらつきは，作業の開始時，あるいはアノテーションの所要時間が長い際に生じやすく，またそれが訓練データの信頼性低下につながる傾向にあること．
- 本報告で示す実験にて，全てのワーカにおいて，disgust と anger のアノテーションにブレが生じており，α値も低いことから，この 2 項目はアノテーションが難しいタスクであること．
- 本来の想定とは異なるタグが付けられる画像について，ワーカごとに傾向があること．

今後の課題として，本研究で得られた考察結果をもとに，訓練データの信頼性を向上する手法を確立することがあげられる．本研究では，α値を利用することで，アノテーション作業の経過時間や 1 枚の画像にアノテーションする所要時間がデータの信頼性にどのような影響を及ぼしているのかを分析した．また，可視化結果からアノテーションが難しいと推察される画像群を見つけた．信頼値が低いと算出された画像群や，アノテーションが難しいと分析された画像群を優先的に訂正することで，どれほど訓練データの信頼性が向上するのかを検証したい．そして，算出された信頼値に応じて重みをつけてアノテーションの平均値を計算して訓練データを作成した場合と，単純平均を用いて訓練データを作成した場合とで，どれほどデータの品質に差が出るのかを検証したい．加えて，アノテーションの傾向をより考察しやすくする可視化手法も今後模索したい．

また，本稿では 3 人であったワーカの人数を増やして冗長性をあげることで，さらなる観察をしたいと考えている．

謝　辞

本研究の一部は，日本学術振興会科学研究費補助金の助成に関するものです．
ユーザーテストにご協力をいただいた皆様に感謝いたします．

参　考　文　献

[1] 寺本拓真, “精度の高い教師データを作成する仕組みと運用”, Speaker Deck. 2019. https://speakerdeck.com/abeja/afalsetesiyondejing-du-falsegao-ijiao-shi-detawozuo-cheng-suruwei-nibi-yao-nashi-zu-mi, (参照 2021-05-29)

[2] 村上綾菜,伊藤貴之,”機械学習の訓練データの注釈作業のヒートマップによる可視化”，映像情報メディア学会技術報告, 44(10), 253-256, 2020.

[3] Barrett, L. F., Gross, J., Christensen, T. C., and Benvenuto, M., Knowing what you’re feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. Cognition and Emotion, 15, 713-724, 2001.

[4] 丹羽彩奈, 松田寛, “個人間の感情理解の揺れを考慮した感情分析に向けた試み”，人工知能学会全国大会論文集, 2021.

[5] Dawid, A. P. and Skene, A. M., Maximum likelihood estimation of observer error-rates using the EM algorithm, J. Royal Statistical Society, Series C (Applied Statics), 28(1), 20-28, 1979.

[6] 光田航, 飯田龍, 徳永健伸, ”アノテーションとアノテーション作業者の信頼性推定”, 言語処理学会第 21 回年次大会, 553-556, 2015.

[7] 駒谷和範, 岡田将吾, 西本遥人, 荒木雅弘, 中野幹生, ”配布可能なマルチモーダル対話データの収集とアノテーション不一致傾向の分析”, 人工知能学会第 84 回言語・音声理解と対話処理研究会, 45-50, 2018.

[8] Fleiss, J. L., Levin, B., and Paik, M. C., 1981. The measurement of interrater agreement. Statistical methods for rates and proportions 2, 212-236, 22-23.

[9] Krippendorff, K. “Computing Krippendorff’s alpha-reliability”, https://repository.upenn.edu/asc_papers/43/,2011.

[10] 稲垣和人, 吉川大弘, 古橋武, “スペクトラルクラスタリングを用いたアンケートデータ解析に関する一検討”, 情報処理学会第 75 回全国大会, 1-669-670, 2013.

[11] Ebner, N., Riediger, M., & Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. Behavior research Methods, 42, 351-362. doi:10.3758/BRM.42.1.351.

[12] 児玉貴志, 田中リベカ, 黒橋禎夫, “映画推薦対話を具体例とした話者内部状態の推定による対話管理”, 自然言語処理, 28(1), 104-135, 021.

[13] Landis, J. R., Koch, G.G., “The measurement of observer agreement for categorical data. “ Biometrics. 33, 159-174, 1977.

[14] 対馬栄輝, ”信頼性指標としての級内相関係数”, http://www.hs.hirosaki-u.ac.jp/pteiki/research/stat/icc.pdf.

[15] Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., NbClust: an R package for determining the relevant number of clusters in a data set. J Stat Softw. 2014;61:1–36.

[16] Haziza, D., Rapin, J., Synnaeve, G., “Hiplot, interactive high-dimensionnality plots”, https://github.com/facebookresearch/hiplot, 2020.