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 (指導教員：伊藤貴之)

1　はじめに

機械学習の精度や信頼性を向上させるためには，上流工程であるアノテーションの精度をあげることが重要である．しかし人手で行われるアノテーションはワーカの能力や経験などの属人的な理由からデータにばらつきやブレが生じる．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てしまう．特に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このばらつきは大きくなりやすい．そこで本研究では主観によってアノテーションされたデータの可視化・信頼性の評価を行うことで，ワーカーごとのアノテーション傾向を観察し，データの品質を向上させる手法を模索する．

2　関連研究

本章では 2 種類に分けて関連研究を紹介する．1 つ目はデータの信頼性をデータの一致率などから評価した研究である．駒谷ら [1] は，アノ テーションの信頼性を評価する際に Fleiss’s κ ， Krippendorff’s αの 2 つの指標を導入して比較し，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．そのほかにも信頼性を評価する際にワーカの視線情報を用いた研究もある．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．2 つ目はアノテーションデータを可視化し分析している研究である．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 で可視化することでワーカごとにアノテーション傾向が存在することを示した．また，駒谷ら [1] はアノテーション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．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やデータの信頼性と可視化結果の関係などまで含めて包括的に取り組んではいない．

3　提案手法

3.1　使用したデータ

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[3] を使用した．このデータセットは 57 人の若年者，56 人の中年者，57 人の高齢者の計 171 名が参加したもので，6 つの表情 (happiness, disgust，anger，neutrality，sadness and fear) を提供している．このデータベースに含まれる顔画像 977 枚を対象として，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．

3.2　信頼性の評価手法

本研究では，ワーカの主観によってアノテーションされるタスクを採用しているのでワーカ間の回答は一致するとは限らない．そこでワーカの回答の一致率からデータの信頼性を評価する．その評価手法として，本研究では 2 人以上のワーカの一致度を計算する指標である Krippendorff’s α を用いる． α 値は，-1 ≤ α ≤ 1 を取る．先行研究 [1] で報告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，本研究のようなワーカの主観に依存するタスクではアノテーションの一致率が低くなる傾向にあり α の値は 0.4 程度が妥当であるとされている．

表 1: 評価項目

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

3.3　時間とデータ品質の関係性の分析手法

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データの品質に関係するのかを把握するため可視化を行い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. ワーカのアノテーション作業中にログを取得することで得た顔画像 1 枚あたりのアノテーション所要時間と経過枚数のデータを利用する．このデータの前処理として標準化処理を施している．
2. 標準化を施したデータに対して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ている．
3. 最適クラスタ数を決定する．
4. クラスタに色を割り当て散布図で可視化する．
5. Krippendorff’s α 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 α 値を照合する．これによりアノテーションの時間とデータの品質の関係を考察する．

3.4　ワーカのアノテーション傾向可視化手法

本研究では，多次元データである訓練データを以下の 2 種類の方法で可視化した．

3.4.1　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．掲載しきれない具体的な可視化の結果や分析は他の文献 [4] に掲載している．

4.1　分析結果 1-時間変化の関係-

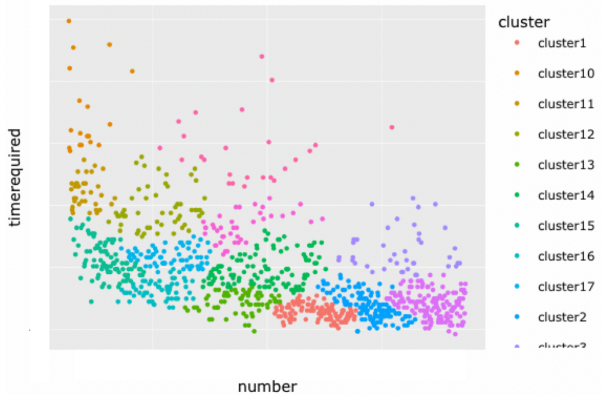

図 1: 経過枚数と所要時間を表す散布図.

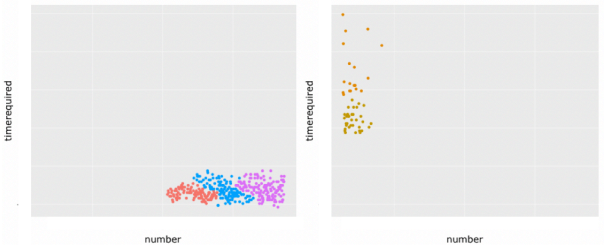

図 2: (左) α 値が高いクラスタ (右) α 値が低いクラスタ.

3.3 節で述べた手順に沿って，Rstudio の shiny パッケージで可視化した結果を述べる．作成した散布図を図 1，図 2 に示す．点 1 個が画像 1 枚に相当しており，横軸はアノテーションした画像の経過枚数，縦軸は 1 枚の画像をアノテーションするのに要した時間を表している．また 3.3 節の手法で決定した 17 個のクラスタに固有の色を割り当てている．図 2 はクラスタごとに α 値を計算した結果，それぞれ α 値が高く出たクラスタと低く出たクラスタを示したものである．図 2 から α 値が高いクラスタはアノテーション作業の終盤で 1 枚あたりにかかる所要時間が短いのに対して，α 値が低いクラスタはアノテーション作業開始時で所要時間が長いクラスタであることがわかる．ここからデータののブレやばらつきは作業の開始時や所要時間が長い際に生じやすい可能性があると分析した．

4.2　分析結果 2-ワーカのアノテーション傾向可視化-

4.2.1　主成分分析での可視化

Rstudio の shiny パッケージにおいて主成分分析を利用して次元削減を行い，各ワーカの多次元の訓練データを可視化した．可視化結果から 3 人とも anger と disgust のアノテーションにブレが生じていることを発見した．また，ワーカごと項目ごとに算出した α 値においてもこの 2 項目を観察すると disgust と anger は 3 人とも α 値が低く出ておりデータにばらつきが生じていた．これらの結果から anger と disgust のアノテーションタスクは難しいと推測することができる．

4.2.2　平行座標プロットでの可視化

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 というライブラリで訓練データを可視化した．その結果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可能性を示唆した．

5　まとめと今後の展望

本研究では，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手法を提案し，信頼性の高いアノテーションを実現する手法について議論した．その結果，1 枚の画像あたりのアノテーションの所要時間や経過時間がデータのブレに関係していることやアノテーションが難しい項目を発見した．

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挙げられる．可視化結果や α 値から信頼性が低いデータであると考察された画像群を優先的にアノテーション訂正することや α 値に応じて重みつけを行いデータ作成することでどれほど信頼性が向上するのかを確かめたい．また，ワーカの傾向をより深く考察できるような可視化手法を検討したいと考えている．

謝辞

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．

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．

参考文献

[1] 駒谷和範，岡田省吾，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.

[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] Ebner, N., Riediger, M., and 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.

[4] 三浦梨花，栃木彩実，伊藤貴之，“主観を要するアノテーションタスクの観察と可視化”，DEIM，2022.