基準例を用いた主観評価からのバイアス軽減に向けて

Toward a Design for Debiasing Subjective Assessment with Criteria

大橋 英明 † 清水 敏之 † 吉川 正俊 †
Hideaki Ohashi Toshiyuki Shimizu Masatoshi Yoshikawa

概 要

5 段階評価のような主観評価は不定形な成果物を評価する際に有効であるが，多くのバイアスを含み得るため，信頼性の高い評価値の推定を目標とした研究が盛んに行われている．本研究は評価者が少人数しかおらず，複数の評価結果を統合して信頼性の高い評価値を得ることが困難な状況に焦点を当て，各々の評価者からバイアスを軽減させるための枠組みを提案する．具体的には，同一課題（例：「優れたデザインを持つウェブサイトの例を挙げ，説明せよ」）における過去の評価対象集合（例：前年度に提出されたレポート集合）から評価基準となる対象を抽出し，評価者に参照させることで，評価者が持つバイアスを軽減させるような枠組みとなっている．本論文では評価基準となる対象を基準例と呼び，基準例が満たすべき条件を整理した．今後，基準例を抽出する手法を開発した後に，基準例の有用性を示すための実験を行う予定である．

1. はじめに

「優れたデザインを持つウェブサイトの例を挙げ，説明せよ」「Ruby on Rails を用いてアプリケーションを作成せよ」といった曖昧な課題に対する不定形な成果物を評価する際，評価者の主観評価は欠かせない．しかし，主観評価は個人の裁量に任されている部分が大きいため，多くのバイアスを含み得る．そのため，複数の評価者による評価結果を統合して信頼性の高い評価を得る手法を提案した研究 [1, 2, 8] や，評価者から正確な評価を誘引するメカニズムを提案した研究 [5, 4, 3] など、評価精度を向上させるための研究が盛んに行われている．前者は十分な数の評価者を前提としたものが多いため，評価者が少数である場合には後者のようなアプローチが不可欠である．本研究は少人数での評価において，次に述べるルーブリックのような評価のための客観的基準が有用であると考えた．

近年，教育の現場において，課題の評価基準を体系的に定義するルーブリックと呼ばれる枠組みが用いられている．ルーブリックはある課題を複数の要素に分け，要素ごとに評価基準を満たす条件を説明したものである（図 1 参照）．評価者である教員はルーブリックを参照することで，効果的・効率的に成績を評価できる [6]．ルーブリックは教育分野以外での主観評価に対しても有用だと考えられるが，評価課題ごとに最適な評価基準を作成するには専門家による入念な考察が欠かせず，非専門家の判断だけで作成することは困難である．そこで，本研究は自然言語での評価基準の説明を用いず，各評価基準となる評価対象例（**基準例**と呼ぶ）を用いて主観評価を

評価課題

	評価尺度1（良い）	評価尺度2（普通）	評価尺度3（悪い）
評価要素1	評価基準1-1	評価基準1-2	評価基準1-3
評価要素2	評価基準2-1	評価基準2-2	評価基準2-3
評価要素3	評価基準3-1	評価基準3-2	評価基準3-3

図 1: 基本的なルーブリックの表

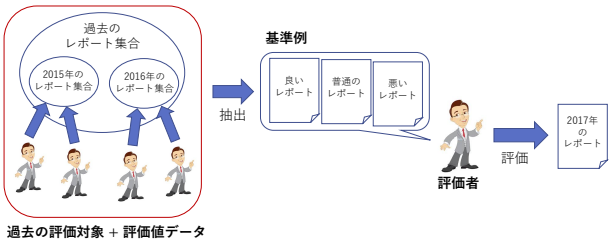

図 2: 枠組みの概要例

補助するような枠組みを提案する．ただし，過去に評価された同様の評価課題における評価対象の集合から，複数の評価者によるリッカート尺度での評価値データを利用して，基準例を抽出する枠組みとなっている．例えば，ある課題に対するレポートの成績評価をする際，同一課題に対して過去に提出されたレポートの中から，複数の TA による過去のレポートの採点結果を利用して基準例が抽出されるものとする（図 2 参照．ただし，図 2 では レポートの評価要素が一種類であると仮定している）．

本研究は基準例の提示により，評価者から評価傾向に関するバイアスを軽減させることを目的とする．評価傾向の例として，対象を高く評価する傾向（寛大化傾向）や低く評価する傾向（厳格化傾向），評価が中心に偏る傾向（中心化傾向）などが存在する [7]．例えば，1 の評

† 京都大学大学院情報学研究科, Graduate School of Informatics, Kyoto University

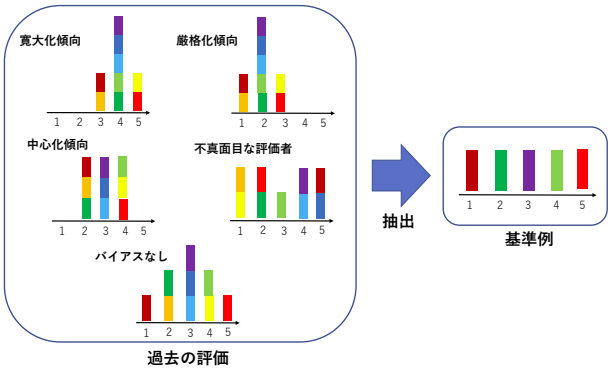

図 3: 基準例の抽出過程

価が最も低く，5 の評価が最も高い，5 段階のリカート尺度を用いた主観評価を想定すると，寛大化傾向を持つ評価者は 3-5 といった上位に偏った評価を，厳格化傾向を持つ評価者は 1-3 といった下位に偏った評価を，中心化傾向を持つ評価者は 2-4 といった中心に偏った評価を行うと考えられる．これらのバイアスを持つ評価者に対して，基準例を提示し，広がりを持った評価を付けるよう促すことは有用だと考えられる．

本研究は上述の目的に合致した基準例の抽出を目指す．ただし，過去の評価対象がそもそもバイアスを持った評価者によって評価されている場合が多いと考えられる（図 3 参照．ただし，図 3 の多様な色を持つ長方形はそれぞれ異なる評価対象を表しており，各長方形の下に記述された数字がその評価対象に対する評価値を表している．例えば，寛大化傾向を持つ評価者は二つの評価対象に対して 3 を，五つの評価対象に対して 4 を，二つの評価対象に対して 5 の評価を下している）．よって，過去の評価値で単純な多数決を取り基準例を決定するのではなく，過去の評価者の傾向を見積もった上で，基準例を抽出することが効果的であると考えられる．そこで，我々は多数の評価者の評価値データから評価者の特徴と真の評価値を推定する既存研究 [1, 2, 8] を参考に，基準例抽出手法を提案する予定である．

本論文の構成を以下に示す．第 2 節では，主観評価に関する関連研究の紹介を行う．第 3 節では，基準例が持つべき特徴について整理する．第 4 節では，今後について述べる．

2. 関連研究

2.1 複数の評価結果から評価値を推定する研究

Dawid らは評価者の能力を考慮しながら，評価対象の真の評価値を推定する基礎的な枠組みを提案している [1]．また，この研究の応用として，時間経過とともに変化する評価者の能力を考慮した研究 [2] や，特定の評価対象集合に対して評価を行う際に発生するバイアスを考慮した研究 [8] などが存在する．本研究はこれらの既存研究を参考に，基準例抽出手法を開発する予定である．

2.2 評価者から正確な評価を誘引する研究

Prelec らは評価者に対して他の評価者の評価結果を予測させ，その値を用いることでより正確な情報を引き出すベイジアン自白剤と呼ばれる枠組みを提案した [5]．また，複数のインセンティブを設計し，それらの有効性を比較した研究 [4, 3] も存在する．本研究は客観的な評価基準の提示によって，評価者から正確な評価を誘引できると考えた．

2.3 ルーブリック

ルーブリックは教育評価のための枠組みの一つであるが,「事前に作成したルーブリックを学生に公開することで，課題の質の向上が見込まれる」「学生とともにルーブリックを作成することで，学生の課題への理解が深まる」といった副次的な効果についても多く研究されている [6]．本研究ではこのような副次的な効果に着目せず，ルーブリックを評価のための枠組みとしてのみ捉える．

3. 基準例について

本研究は過去の評価対象の中から過去の評価値データを用いて基準例を抽出し，基準例を評価者に参照させることで評価者の評価傾向に関するバイアスの軽減を行う枠組みを考える．基準例が持つべき条件は以下のように整理できると考えた．

1. 評価値の分散が小さいこと
2. 特異な内容を含んでいないこと
3. 簡潔であること

一つ目は各基準における評価対象の中から特に信頼度が高いもの（当該基準である可能性が高いもの）を基準例として抽出するための条件である．二つ目は基準例に含まれる特異な内容によって基準として用いることが困難な状況を避けるための条件である．三つ目は評価者が基準例を参照するためのコストをなるべく小さくするための条件である．

4. まとめ・今後について

本研究は基準例の提示によって主観評価からバイアスを軽減させる枠組みを提案し，基準例が持つべき条件について整理した．今後は 3 節で上げた条件を満たす基準例を抽出する手法を開発し，実験によって有用性を示したい．

参考文献

[1] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pp. 20–28, 1979.

[2] Pinar Donmez, Jaime Carbonell, and Jeff Schneider. A probabilistic framework to learn from multiple annotators with time-varying accuracy. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pp. 826–837, 2010.

[3] Boi Faltings, Radu Jurca, Pearl Pu, and Bao Duy Tran. Incentives to counter bias in human computation. In *Second AAAI conference on human computation and crowdsourcing*, 2014.

[4] Christopher Harris. You're hired! an examination of crowdsourcing incentive models in human resource tasks. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 15–18, 2011.

[5] Dražen Prelec. A bayesian truth serum for subjective data. *science*, Vol. 306, No. 5695, pp. 462–466, 2004.

[6] Dannelle D Stevens and Antonia J Levi. *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*. Stylus Publishing, LLC, 2013.

[7] Edward W Wolfe. Identifying rater effects using latent trait models. *Psychology Science*, Vol. 46, pp. 35–51, 2004.

[8] Honglei Zhuang, Aditya Parameswaran, Dan Roth, and Jiawei Han. Debiasing crowdsourced batches. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1593–1602, 2015.