# 異なる評価基準が混在したリッカート尺度の正規化手法

川上　大凱† 　鈴木　優†

† 岐阜大学大学院自然科学技術研究科知能理工学専攻　501-1193 岐阜県岐阜市柳戸 1-1
E-mail: †kawakami.taiga.b8@s.gifu-u.ac.jp, ††suzuki.yu.r4@f.gifu-u.ac.jp

**あらまし**　本研究の目的は，リッカート尺度に基づく主観評価に内在する個人差の解消である．本研究の目的を達成するため，我々はリッカート尺度に基づく主観評価を多数の評価者による平均的な判断基準に対応するスケールへと正規化する手法を提案する．主観評価には，同一の評価値であっても評価者ごとに判断基準が異なるという問題がある．特に，実際に付与された満足度と評価理由を示すレビューから推測される満足度が大きく異なる例外的なレビューケースにおいてはその傾向が顕著である．このことから我々は，例外的なレビューケースを主観評価に個人差が内在する要因の一つであると考えた．よって本稿において我々は，既存研究ではノイズとして扱われてしまう例外的なレビューケースの存在を構造的な情報として保持することが可能な混同行列を用いてレビュアごとの評価傾向を表現する．実験では，本提案手法が個人差解消の正規化タスクにおいて有効かどうかを検証した．実験結果より，意味解釈の個人差を構造として扱うことによって，既存研究ではノイズとして扱われていた個人差が体系的に補正可能であることが示された．一方で，混同行列における個人差を表す構造によって補正効果が変わることを確認した．

**キーワード**　レビュー，評価傾向，LLM，Zero-Shot 推論，ニューラルネットワーク，リッカート尺度，正規化

## 1　はじめに

本研究で我々は，リッカート尺度に基づく主観評価を多数の評価者による平均的な判断基準に対応するスケールへと正規化する手法を提案する．本提案手法により，リッカート尺度に基づく主観評価における意味解釈の個人差を解消することが可能となる．これにより，本提案手法は評価値の直接的な比較や集計から導かれる分析結果の安定した解釈を可能とする．

リッカート尺度は，EC サイトに投稿されるサービス評価やアンケート調査などで用いられる満足/不満足といった主観評価を表す順序尺度の一つである．リッカート尺度は，企業のマーケティング分析や商品改善のための重要なデータ資源としても活用されている．またマーケティングだけでなく，心理学のような主観評価を定量的に利用する研究分野においてもリッカート尺度は多く活用されている．しかし，リッカート尺度に基づく主観評価には大きな問題がある．その一つが，主観評価に内在する個人差である．主観評価は，評価者の評価基準や考え方によって個人差が生まれる．例えば評価が全体的に厳しい評価者と寛容な評価者では，同一の評価値が必ずしも同一の意味を内包しない．そのためリッカート尺度に基づく主観評価における個人差は，評価値の直接的な比較や集計から導かれる分析結果の解釈を不安定なものにしてしまう．したがって，本研究は主観評価における意味解釈の個人差を解消することを目的とする．

我々はこの問題を解消するために，リッカート尺度に基づく主観評価を意味解釈の個人差が内在する主観尺度から意味解釈の個人差が存在しない参照尺度へと変換する必要があると考えた．我々は，評価値とは独立に存在するものではなく評価者が持つ評価理由に基づいて付与された結果であるという立場に立つ．この立場に基づく場合，主観評価に内在する個人差は評価理由と評価値における対応関係の違いとして捉えることが可能となる．我々は評価理由と評価値における対応関係の違いを基に評価値を再解釈することによって，評価者ごとに異なる意味解釈を持つ主観的な尺度を意味解釈の個人差が存在しない参照尺度へと変換することが可能であると考えた．

本研究において我々は，リッカート尺度に基づく主観評価に内在する個人差を解消するために，評価者の評価傾向を用いる．本稿において我々は，評価者の評価傾向を評価理由と評価値における対応関係と定義する．我々は評価者の評価傾向を定量的に表現するために，評価者が付与した評価値と評価理由を表すレビューから推測される評価値の共起頻度を利用する．既存の Z-score 正規化は，平均や分散といった統計量のみに基づいて個人差を補正する．そのため，評価者の評価傾向や考え方に関する特徴的な情報は失われてしまう．一方で既存の IRT 系モデルによる正規化は，評価者の評価傾向を能力や厳格さといったスカラ値を用いて統計確率的に推定する．そのため，多数の評価者の平均的な評価基準に基づいた個人差の補正が可能となる．しかし，IRT 系モデルは評価者の評価傾向を単一のスカラ値に落とし込んでいるため，例外的なレビューケースがノイズとして平均化されてしまう．例外的なレビューケースとは，レビュー内では不満点や改善点といったネガティブな言及がされている一方でポジティブな評価値が付与されるといったレビューと評価が大きく異なるケースのことである．我々は，例外的なレビューケースを主観評価に個人差が内在する要因の一つであると考えた．したがって，例外的なレビューケースがノイズとして平均化されてしまう場合，評価者の正確な評価基準をもとに個人差を補正しているとは言い難い．一方で，評価者が付与した評価値と評価理由を表すレビューから推測される評

価値の共起頻度は例外的なレビューケースを共起パターンとして保持することが可能である．

本稿において我々は，本手法の検証対象として EC サイトのレビューデータを扱う．そのため我々はレビュアの評価傾向を，共起頻度が集約されたデータであるレビュア単位の満足度予測における混同行列を用いて表現する．混同行列は，そのレビュアの感じた実際の満足度がどの程度レビューに内包されているのかを要素の大きさや位置で表現可能である．本提案手法の有効性を検証するため，混同行列とレビュアが付与した満足度を入力とし，意味解釈の個人差が存在しない参照尺度へと正規化した満足度を出力とするニューラルネットワークを構築する．本稿において我々は，意味解釈の個人差が存在しない参照尺度，つまり多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度を LLM による Zero-Shot 推論の予測値で定義した．以降，多数のレビュアに共通する平均的な評価傾向を近似的に表現する参照尺度を近似参照尺度と呼称する．実験結果より，提案手法が一部の条件下で主観評価に内在する個人差の解消に有効であることを確認した．

## 2 関 連 研 究

Yeşilçınar ら [1] は，リッカート尺度に相当する段階的な評価に基づくピア評価を対象とし，評価環境が評価者バイアスに与える影響を定量的に明らかにすることを目的とした．ピア評価とは，同一または類似した専門分野に属する評価者同士が互いの成果物を評価し合う活動を指す．評価環境とは，対面やオンライン，匿名といった評価時の環境を指す．評価者バイアスとは，段階評価において中央付近の評価を付与しやすいという中央化効果 [2] や，相手の学歴や自分との関係性といった他の目立つ情報によって正確な評価が歪められてしまうハロー効果 [3] などを指す．Yeşilçınar らは，評価者同士の関係性と評価環境に応じて主観に基づいたピア評価の信頼性が変動すると考えた．同研究において Yeşilçınar らは，IRT を拡張したモデルである MFRM を用いた．MFRM は，評価者の厳しさや評価環境の影響をスカラ値の潜在パラメータとして推定する．そのため，各レビュー単位における評価理由と評価値の乖離は誤差項として吸収される．つまり，MFRM は評価理由と評価値が大きく乖離する例外的なレビューケースを構造的に保持することが困難なモデルである．また，MFRM はレビュー内容そのものをモデル化の対象としていないことも特徴である．実験では学生同士のピア評価および教員による評価を比較し，評価環境を含む誤差要因のうちどの要因が評価の信頼性を損なうのか定量的に調査した．実験結果より，学生同士の評価においては匿名性の有無によって評価信頼性に有意な差が生じることを示した．また，教員による評価では匿名評価の有無によって評価信頼性に有意な差は確認できなかった．

Anjaria ら [4] は，リッカート尺度に代表される主観評価に内在する個人差から生まれた尺度の不確実性や曖昧さを解決することを目指した．その際 Anjaria らは，z-number [5] を用いることによって段階的な評価尺度における曖昧さを明示的に取り扱うことが可能であると考えた．この考えを基に，Anjaria らは z-number を用いた主観評価の表現手法を提案した．z-number は，評価値に対応する制約とその評価に対する信頼度をファジィ数として表現する枠組みである．制約は，段階的な尺度の各項目において回答がどのように分布しているのかを表す．信頼度は，評価者がどれだけ確信を持って評価したのかを表す．ファジィ数とは，大きさや信頼性といった概念的な値を離散値ではなくどの程度大きいのか，信頼できるのかといった連続値として表現した数を指す．つまり z-number に基づく表現は，評価者の評価傾向を評価値の集合および信頼度のスカラ値を用いて集約した要約的な情報で表現する．そのため，個々のレビューにおける評価理由と評価値の対応関係を構造的に保持することは困難である．また，前述した MFRM と同様に例外的なレビューケースは曖昧さの一部として吸収される．実験では，講義に対する教師と学生の認識の違いを z-number を用いて分析した．実験結果より，教師は学生と比較して高い信頼度を持って主観評価を行っていることを確認した．また学生が否定評価をしている場合は信頼性が低い，つまり反対の意味だけではなく理解ができていないという状態であることを確認した．

Wang ら [6] は，レビューに加えてユーザおよび製品固有のコンテキスト情報を潜在表現として統合することにより，評価値予測精度の向上を目指した．コンテキスト情報とはユーザや製品に固有の特徴を表すものである．同時に，コンテキスト情報とはレビューや製品説明に含まれる感情表現が各ユーザや各製品で異なる影響を持つことを潜在的に表現したものである．同研究では，ユーザや製品ごとの評価の偏りはモデル内部の潜在変数として暗黙的に表現される．つまり，コンテキスト情報は Wang らの研究における主観評価の個人差であり，明示的に設定されたパラメータではない．実験では，6 つのベースライン手法とユーザや製品の固有特徴をレビューに統合した提案手法を比較した．実験結果より，コンテキスト情報を統合することが評価値予測精度向上の観点で有効であることを確認した．

坂本ら [7] は，レビューからレビュアの評価値を予測するモデルを構築し，その予測誤差を用いてユーザの嗜好に適合したレビューを抽出する手法を提案した．この研究は，レビュアが付与したアイテムへの評価値とレビューからモデルが推測したアイテムへの評価値間の差が小さいものを，レビュアの嗜好に合うレビューとしている．そのため，坂本らの研究において評価値予測はレビュアの嗜好を含むレビューを識別するための手段として位置づけられている．実験では，実際の評価値と推測の評価値に差があるものは具体的な内容のないレビュー，もしくはユーザの嗜好のレビューであるという仮定を検証するために，Fine-Tuning を活用した BERT [8] モデルによる評価値予測を実施した．実験結果より，提案手法によって具体的な内容のある役に立つレビューが抽出できることを確認した．

## 3 提案手法

本研究の目的は，リッカート尺度に基づく主観評価に内在する意味解釈の個人差を解消することである．我々は本研究の目的を達成するために，レビュアの評価傾向を利用した意味解釈の個人差が内在しない近似参照尺度への正規化手法を提案する．

既存研究における評価者，つまりレビュアの評価基準は，2 章で述べたように集合やスカラで表される．しかし既存研究で用いられる方法を正規化に適用する場合，例外的なレビューケースにおける主観評価はノイズとして平均化されてしまう．例外的なレビューケースの一例を図 1 に示す．例外的なレビューケースにおける主観評価がノイズとして扱われてしまう場合，我々はレビュアの正確な評価傾向をもとに個人差を補正しているとは言い難いと考えた．本研究はこの点に着目し，主観評価に内在する個人差を構造として扱うことによって評価値に対する意味的正規化を行う点に新規性がある．特に本研究は既存研究と異なり，例外的なレビューケースを個人差の解消のための重要な要素であると位置付ける．つまり，本研究は例外的なレビューケースの存在をノイズとして平均化するのではなく，例外的なレビューケースの存在を構造情報として維持したまま主観評価を扱う．本稿において我々は，レビュアの評価傾向をレビュアが付与した満足度と評価理由を表すレビューから LLM の Zero-Shot 推論によって推測される満足度の共起頻度，つまりレビュア単位の Zero-Shot 満足度予測によって得られる混同行列として定義する．詳細は 3.2 節で述べる．また本稿において我々は，2 章で述べた既存研究と異なり，個人差を補正するための正規化タスクとして満足度予測を利用する．その際，我々は正解データである正規化後の意味解釈の個人差が存在しない近似参照尺度に基づく評価値を LLM による Zero-Shot 推論の予測値で定義する．詳細は 3.3 節で述べる．

### 3.1 提案手法概要

提案手法は，図 2 に示すような 2 つの Step に分けられる．

Step 1　混同行列を用いた評価傾向の用意
Step 2　ニューラルネットワークを用いた主観尺度の正規化

レビュアの集合を $U = \{u_0, u_1, \cdots, u_i, \cdots, u_n\}$，レビューの集合を $R(u_i) = \{r_0, r_1, \cdots, r_j, \cdots, r_m\}$ とする．この時，$0 \leqq i \leqq n$，$0 \leqq j \leqq m$ である．レビュー $r_j$ にはそのレビューに付与された評価値 $y_j$ が付随する．評価値カテゴリ集合を $Y$ とする時，評価値 $y_j$ は $y_j \in Y = \{1, 2, \ldots, k\}$，$k \in \mathbb{N}$ である．したがって，あるレビュア $u_i$ が持つデータ集合は $D(u_i) = \{(r_0, y_0), (r_1, y_1), \cdots, (r_j, y_j), \cdots, (r_m, y_m)\}$ で表す．また，サンプリングした複数のレビュア $U_s = \{u_0^s, u_1^s, \cdots\}$ が持つデータ集合は $D(U_s) = \{D(u_0^s), D(u_1^s), \cdots\}$ で表す．この時，$|D(u_i)| \geqq 1$, $|D(U_s)| \geqq 1$ である．評価傾向は $|D(u_i)|$ や $|D(U_s)|$ が大きいほど明瞭に表現することが可能となる．

Step 1 で我々はレビュアごとの評価傾向を用意する．Step 2 で評価傾向と実際に付与された評価値をもとに，ニューラルネットワークを用いた評価値の正規化を行う．詳細は 3.2 節と 3.3 節で述べる．

### 3.2 評価傾向の用意 (Step 1)

本稿において我々は，主観評価に内在する個人差を定量的に解釈するために，レビュアが書いたレビュー $r_j$ とそれに付随して付与された満足度 $y_j$ の対応関係に着目する．対応関係を定量的に示す際，我々はレビュアが持つデータ集合 $D(u_i)$ における満足度間の差ではなく満足度間の共起関係を用いる．我々が満足度間の共起関係を用いる理由は，順序尺度であるリッカート尺度に基づく評価値間の差が必ずしも同一であると言い切れないためである．これは，各評価値が同等の評価幅を以て定義されているわけではないことが原因である． このことから，本稿において我々はデータ集合 $D(u_i)$ の共起構造を集約したデータで

===レビュー===
スタッフには何の不満もない。が、グループ利用者がうるさい．部屋もいつもより汚い。前に来た時はよかったのに。グループを角部屋にするという配慮くらいしてほしい。
===評価値===
レビュアが付与した実際の評価値：5
レビューから推察するレビュアの評価予測値：2

図 1: 例外的なレビューケースの一例

表 1: レビュア $u_i$ に対する満足度予測の混同行列

| 予測満足度 $\hat{y}_j$ | 実際の満足度 $y_j$ | | | |
|---|---|---|---|---|
| | 1 | 2 | $\cdots$ | $k$ |
| 1 | $C_{1,1}^{(u_i)}$ | $C_{1,2}^{(u_i)}$ | $\cdots$ | $C_{1,k}^{(u_i)}$ |
| 2 | $C_{2,1}^{(u_i)}$ | $C_{2,2}^{(u_i)}$ | $\cdots$ | $C_{2,k}^{(u_i)}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $k$ | $C_{k,1}^{(u_i)}$ | $C_{k,2}^{(u_i)}$ | $\cdots$ | $C_{k,k}^{(u_i)}$ |

表 2: 評価傾向の違いを表す混同行列の例

(a) 順当に満足度を反映している場合

| $\hat{y}\backslash y$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 1 | 2 | 0 | 0 | 0 |
| 2 | 0 | 4 | 3 | 0 | 0 |
| 3 | 0 | 0 | 11 | 9 | 0 |
| 4 | 0 | 0 | 0 | 38 | 25 |
| 5 | 0 | 0 | 0 | 0 | 7 |

(b) 評価とレビューが乖離している場合

| $\hat{y}\backslash y$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 2 | 0 |
| 2 | 1 | 2 | 7 | 4 | 0 |
| 3 | 0 | 1 | 4 | 6 | 1 |
| 4 | 0 | 0 | 2 | 6 | 0 |
| 5 | 0 | 0 | 0 | 3 | 0 |

(c) レビュー数が極端に少ない場合

| $\hat{y}\backslash y$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 2 | 0 |
| 4 | 0 | 0 | 0 | 1 | 0 |
| 5 | 0 | 0 | 0 | 0 | 2 |

(d) 極端な判断を行う場合

| $\hat{y}\backslash y$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 4 | 0 | 0 | 0 | 2 |
| 3 | 16 | 1 | 4 | 0 | 4 |
| 4 | 3 | 0 | 0 | 0 | 14 |
| 5 | 0 | 0 | 0 | 0 | 2 |

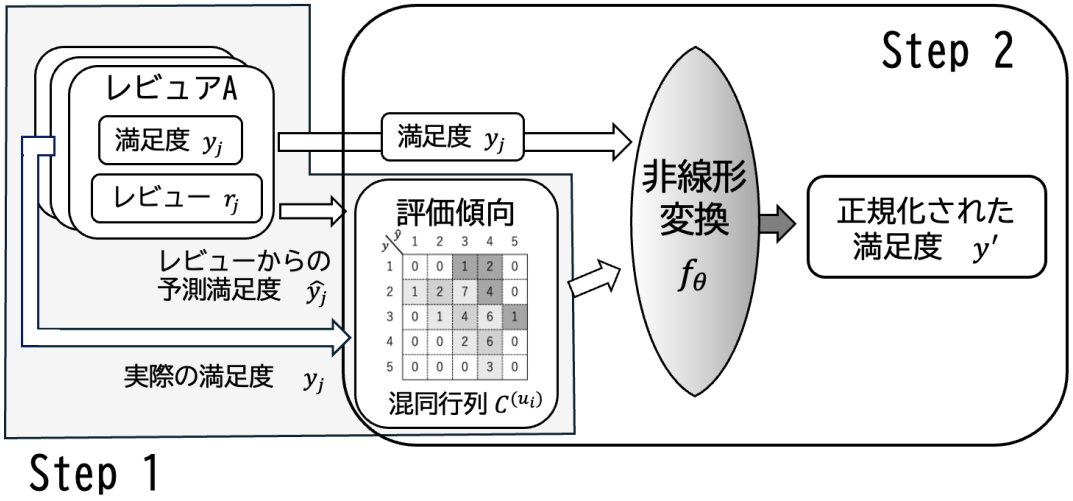


図 2: 提案手法の概要図

あるレビュア単位の満足度予測における混同行列 $C^{(u_i)}$ を，レビュアの評価傾向を表す情報として利用する．混同行列 $C^{(u_i)}$ は表 1 に示すように，評価値カテゴリ集合の要素数 $k$ に依存した行列サイズをもつ．つまり $C^{(u_i)} \in \mathbb{R}^{k\times k}$ であり，混同行列の各要素 $C^{(u_i)}_{p,q}$ は $1 \leqq p \leqq k, 1 \leqq q \leqq k$ である．具体例を表 2a と表 2b，表 2c，表 2d に示す．混同行列 $C^{(u_i)}$ の対角線上にある要素を濃い灰色で示す．それ以外の非ゼロ要素を薄い灰色で示す．表 2a に示すように対角線上に要素が集合している場合，つまり $C^{(u_i)}_{p,q}$ における $p = q$ の位置に要素が集合している時，そのレビュアはレビューを順当に満足度へと反映している．反対に表 2b に示すように，対角線上に要素が集合せず要素がばらついている場合，そのレビュアは言及した要素とは関係のない評価を付与している．表 2b における，$C^{(u_i)}_{1,4}$ や $C^{(u_i)}_{3,5}$ が例外的なレビューケースに該当する要素となる．また表 2c や表 2d に示すように，混同行列の要素が少ない場合や極端な評価が行われる場合も存在する．加えて 2 章で述べた既存研究のように，一次元のベクトルやスカラ値は隣接した評価値間に共通するポジティブ/ネガティブといった感情情報が同一であるか否かを表現できない．したがって，我々は近しい評価値間の意味的な類似性が失われてしまうと考えた．一方で混同行列は，隣接した評価値間に共有される感情情報を実際に付与された満足度 $y_j$ と推測された満足度 $\hat{y_j}$ の対応関係である要素の位置やばらつきによって構造的に保持することが可能となる．以上の事実より，我々は混同行列 $C^{(u_i)}$ が主観評価に内在する個人差を定量的に表現していると考えた．本稿において我々は，$C^{(u_i)}_{p,q}$ における $p = q$ の位置に要素が集合しているレビュアを多数のレビュアが持つ平均的な判断基準に基づいた評価傾向を持つレビュアだと定義する．また本稿において我々は，混同行列 $C^{(u_i)}$ において対角線上に要素が集合せず要素がばらついているレビュアを特殊な評価傾向を持つレビュア，つまり個人差が内在する主観評価を行うレビュアだと定義する．

本稿において我々は混同行列 $C^{(u_i)}$ を計算するために，実際に付与された満足度 $y_j$ と Zero-Shot 推論によって推測された満足度 $\hat{y_j}$ の共起頻度を用いる．共起頻度とは，特定の事象が他の事象と一緒に出現する回数やその傾向を数値化したものである．Zero-Shot 推論とは，LLM の判断時に参考となるデータをプロンプト上で与えないまま，推論したい対象とタスク指示のみを与える推論方法である．つまり本稿における共起頻度は，レビュアが実際に付与した満足度 $y_j$ とレビューのみから推測される満足度 $\hat{y_j}$ の対応関係をレビュア単位で集計したものである．

我々は評価傾向に内在する個人差を，多数のレビュアが持つ平均的な評価基準からそのレビュアの評価基準がどれほど離れているのかを表す指標だと考える．しかし，レビュアの評価基準が平均的な評価基準からどれほど離れているのかを定量的に解釈するためには，平均的な評価基準を定量的に表現することが必要となる．多数のレビュアにおける平均的な評価基準を定量的に表現する際，我々は全てのレビュアに対して判断基準を推定するための調査を行う必要があると考えた．しかし，実際に同じサービスを受けたレビュア全員に対して調査を行うことは現実的ではない．そこで我々は，事前学習済み LLM を用いて近似参照尺度を定義する．事前学習済み LLM は，多様な文脈や表現を含む大規模コーパスに基づいて学習されている．そのため，事前学習済み LLM は一般的知識や典型的な判断傾向に基づく推論を行う能力を有する．このことから我々は，事前学習済み LLM が近似参照尺度として適切であると考えた．よって本稿において我々は，LLM の持つ一般的知識や典型的な判断傾向を崩さずに判断させるため，事前学習済み LLM による Zero-Shot 推論によって得られる満足度予測値 $\hat{y_j}$ を用いて近似参照尺度を定義する．

レビュアが実際に付与した満足度 $y_j$ は，レビュアの評価傾向に基づいて尺度空間上に直接射影された満足度である．事前学習

済み LLM による Zero-Shot 推論によって得られる満足度予測値 $\hat{y_j}$ は，近似参照尺度に基づいて意味空間上に射影された満足度である．つまり，レビュアが実際に付与した満足度 $y_j$ とレビューから推測される満足度 $\hat{y_j}$ の対応関係を集計した混同行列 $C^{(u_i)}$ は，あるレビューに基づいて意味空間上に射影した満足度に対してそのレビューを書いたレビュアがどの満足度を付与しやすいかという共起構造を表現している．このことから我々は，実際に付与された満足度 $y_j$ と Zero-Shot 推論によって推測された満足度 $\hat{y_j}$ の共起頻度を集計することによって，レビュアの評価傾向が近似参照尺度からどれほど離れているのかを表現することが可能だと考えた．

### 3.3 リッカート尺度の正規化 (Step 2)

我々は，個人差の内在するリッカート尺度に基づく満足度 $y_j$ とレビューを書いたレビュアの混同行列 $C^{(u_i)}$ を用いて，近似参照尺度に基づく満足度 $y'_j$ を射影する．この際，我々はニューラルネットワークを構築し，近似参照尺度に基づく満足度 $y'_j$ への射影，つまり尺度の正規化へと活用する．

ニューラルネットワークの入力 $\mathbf{x}_j^{(u_i)}$ は，正規化対象の満足度 $y_j$ とそのレビューを書いたレビュアの混同行列 $C^{(u_i)}$ を結合した一次元のベクトル $v(C^{(u_i)})$ である．この時，$\mathbf{x}_j^{(u_i)} = [y_j, v(C^{(u_i)})] \in \mathbb{R}^{\dim(\mathbf{x}_j^{(u_i)})}$ である．入力次元数 $\dim(\mathbf{x}_j^{(u_i)})$ は $\dim(\mathbf{x}_j^{(u_i)}) = N(y_j) + |C^{(u_i)}|$，つまり $\dim(\mathbf{x}_j^{(u_i)}) = 1 + k^2$ である．ニューラルネットワークの出力 $y'_j$ は，入力 $\mathbf{x}_j^{(u_i)}$ に基づく射影である．以上より，本研究のタスクにおけるニューラルネットワークの入出力は以下の式で表される．

$$y'_j = f_\theta\left(\mathbf{x}_j^{(u_i)}\right), \quad f_\theta : \mathbb{R}^{\dim(\mathbf{x}_j^{(u_i)})} \to \mathbb{R} \tag{1}$$

この時，$\theta$ は正規化写像 $f_\theta$ を構成するニューラルネットワークの学習可能パラメータである．

本研究のタスクは，回帰タスクとして設計する．分類タスクではなく回帰タスクとした理由は，分類モデルの構造に関係する．分類タスクに用いるモデルは，出力層に評価値カテゴリ集合の数だけノードを持つ．それに伴い，出力は各クラスに属する確率となる．このことから我々は，分類モデルにおける最終的な予測満足度はモデル内部で計算された確率に基づく名義尺度のようなものであると考えた．正規化された後の満足度は近似参照尺度に基づく満足度である．そのため，我々は隣接した満足度評価値を近い満足度評価値だと認識できるよう，順序と距離がない名義尺度でのシステム化は不適切だと考えた．学習時の教師信号は，近似参照尺度として定義した Zero-Shot 推論による満足度予測結果 $\hat{y_j}$ を利用する．前述した通り，事前学習済み LLM は多様な文脈や表現を含む大規模コーパスに基づいて学習されている．そのため，事前学習済み LLM は特定のレビュアの評価傾向に依存しない近似参照尺度を教師信号として与えることが可能となる．

次に，ニューラルネットワークを用いた理由を説明する．我々がニューラルネットワークを選択した理由は，二つある．一つ目は，参照尺度への射影という特殊なタスクへの対応が可能であるためである．自然言語を用いたタスクでは，ニューラルネットワークではなく LLM が用いられることも多い．LLM は多様な自然言語タスクに対して高い汎化性能を示すためである．しかし，LLM は本研究で対象とするような評価傾向とリッカート尺度に基づく主観尺度の対応関係からレビューを基盤とした近似参照尺度への正規化を実施することを目的とした設計をなされていない．また，LLM はプロンプトの質により推論時の振る舞いが変化してしまう．一方で，ニューラルネットワークは一からパラメータを調整することによって，固有の評価傾向に基づく満足度から近似参照尺度に基づく満足度への写像を学習することが可能である．そのため，我々は LLM ではなくニューラルネットワークを用いるべきだと考えた．二つ目は，線形なモデルで特殊な評価傾向に対応することが困難なためである．本タスクにおいて，LLM 以外にも重回帰分析のような線形モデルを用いるという選択肢が存在する．重回帰分析は，各入力特徴が出力に対して独立かつ線形に寄与することを仮定したものである．しかし，本研究で扱う入力は実際に付与された満足度 $y_j$ と，レビュアの評価傾向を表す混同行列 $C^{(u_i)}$ である．つまり，入力と出力の関係は混同行列 $C^{(u_i)}$ による条件付けで変化する．したがって，入力が同一の満足度 $y_j$ であったとしても評価傾向を示す混同行列 $C^{(u_i)}$ が異なる場合は正規化後の満足度 $y'_j$ が変化する可能性がある．この現象は，表 2d のような評価傾向を持つレビュアに顕著である．このことから我々は，本研究のタスクが満足度 $y_j$ と評価傾向 $C^{(u_i)}$ の相互作用を考慮した非線形写像であると考える．非線形写像を表現するためには，特徴間の非線形な関係を学習可能なモデルが必要である．よって我々は，線形モデルを利用するのではなくニューラルネットワークを用いるべきだと考えた．

## 4 実　　験

本実験の目的は，提案手法が主観評価に内在する個人差の解消に有効であるという仮説を検証することである．この実験では，混同行列による評価傾向表現を用いることによって主観評価に内在する個人差の補正とそれに伴う意味空間上での正規化が可能であるかを検証・考察する．具体的には，回帰ニューラルネットワークへ混同行列と正規化対象の満足度を入力し，近似参照尺度として定義した Zero-Shot 推論による満足度予測結果を出力する．実験手順は 3 章で述べた通りである．

### 4.1 実 験 準 備

本実験におけるタスクは，正規化満足度の回帰タスクである．ニューラルネットワークの入力層は 3.3 節で述べたとおり $1+k^2$，つまり 26 次元である．中間層は 13 次元 1 層とした．出力層は回帰タスクであるため，1 次元である．活性化関数は ReLU 関数である．層間は全結合とする．使用するデータは，楽天トラベルデータセット [1]から抽出した 350 人のレビュアにおけるレ

1：楽天グループ株式会社 (2014): 楽天データセット．国立情報学研究所情報学研究データリポジトリ．(データセット)．https://doi.org/10.32130/idr.2.0

表 3: 10 分割交差検証での結果

| Fold | 提案手法 | | | | ベースライン手法 | | | |
|---|---|---|---|---|---|---|---|---|
| | RMSE | MAE | 決定係数 ($R^2$) | Accuracy | RMSE | MAE | 決定係数 ($R^2$) | Accuracy |
| 1 | 0.5626 | 0.4084 | 0.3988 | 0.6687 | 0.7069 | 0.5232 | 0.0508 | 0.5472 |
| 2 | 0.5625 | 0.4047 | 0.3995 | 0.6744 | 0.7001 | 0.5176 | 0.0697 | 0.5481 |
| 3 | 0.5667 | 0.4100 | 0.3905 | 0.6805 | 0.7161 | 0.5297 | 0.0269 | 0.5356 |
| 4 | 0.5753 | 0.4152 | 0.3719 | 0.6722 | 0.7062 | 0.5138 | 0.0534 | 0.5670 |
| 5 | 0.5695 | 0.4184 | 0.3815 | 0.6667 | 0.7151 | 0.5298 | 0.0249 | 0.5424 |
| 6 | 0.5548 | 0.4024 | 0.4130 | 0.6809 | 0.7010 | 0.5209 | 0.0627 | 0.5414 |
| 7 | 0.5646 | 0.4031 | 0.3921 | 0.6809 | 0.7105 | 0.5235 | 0.0372 | 0.5465 |
| 8 | 0.5543 | 0.4016 | 0.4139 | 0.6922 | 0.6897 | 0.5110 | 0.0926 | 0.5571 |
| 9 | 0.5576 | 0.3948 | 0.4069 | 0.6961 | 0.6986 | 0.5131 | 0.0689 | 0.5588 |
| 10 | 0.5402 | 0.3943 | 0.4454 | 0.6841 | 0.6856 | 0.5089 | 0.1067 | 0.5547 |
| **平均** | **0.5608** | **0.4053** | **0.4014** | **0.6797** | **0.7030** | **0.5191** | **0.0594** | **0.5499** |
| **$p$ 値** | **4.986e-14** | **1.008e-12** | **1.145e-13** | **4.490e-11** | - | - | - | - |

ビューと満足度，混同行列である．付与された満足度の割合を維持したまま，訓練用: 検証用: テスト用=8:1:1 としてデータを分割した．割合を維持した理由は，正解ラベルが近似参照尺度に基づく満足度であるためである．訓練用データは 32615 件である．検証用データは 4028 件である．テスト用データは 4027 件である．正解データは，Gemma3:27B[2]による Zero-Shot 推論結果である．量子化は行わない．入力する混同行列の要素は正規化した．正規化した理由は，レビュアごとのレビュー投稿数の違いが特徴量のスケールとして出力に反映されてしまうことを防ぐためである．early stopping は 100 エポックとした．バッチサイズは 32 である．10 分割交差検証を実施した．本実験の目的は主観尺度に内在する個人差の補正であるが，補正の有効性は補正後の予測性能として間接的に評価可能である．そのため，本実験では評価指標として RMSE Loss，決定係数 ($R^2$)，MAE Loss といった回帰用の評価指標に加え，Accuracy を使用した．Accuracy を使用した理由は，本来の満足度が離散値であるためである．Accuracy を計算する際は，得られた連続値の出力を四捨五入して整数とした．

また，本提案手法の比較対象としてベースライン手法を用意する．ベースライン手法は，既存研究を参考にスカラ値を用いた正規化手法である．本実験において我々は，スカラ値としてレビュアごとの平均満足度を利用する．具体的には，レビュアごとの平均満足度と正規化対象の満足度の平均をとったものを正規化後の満足度とする手法をベースライン手法として定義する．ベースラインと提案手法の各精度評価指標は対応のある 2 標本 $t$ 検定を用いて統計的検定を実施した．

### 4.2 結果・考察

実験結果を表 3 に示す．10 分割交差検証の結果，提案手法の決定係数は平均して 0.4014 であり，Accuracy は平均して 0.6797 であった．決定係数はそこまで高い値を取っていない．しかしこの原因は，本来離散値であるカテゴリ変数を回帰で表現したためである．よって我々は，決定係数が低い値を取ること自体は問題ないと考える．一方で，Accuracy は約 7 割ほどである．5 段階の分類タスクにおいてこの結果は悪くないものである．またベースライン手法の決定係数は平均して 0.0594 であり，Accuracy は平均して 0.5499 であった．いずれの $p$ 値も 0.05 を下回ることから，本提案手法がベースラインと比較して有意に精度が高いことがわかる．そしてこの結果は，主観評価に内在する個人差の解消というタスクにおいて本提案手法が有

表 4: 予測誤差が最大のケース（最大予測誤差 = 3.64)

(a) 予測情報

| タイプ | 入力満足度 | 予測満足度 | 正解 | 予測誤差 |
|---|---|---|---|---|
| 予測誤差 (大) | 1.0 | 1.36 | 5.0 | 3.64 |
| 予測誤差 (小) | 4.0 | 3.98 | 4.0 | 0.015 |

(b) 混同行列の全要素

| $\hat{y}\backslash y$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0.05 |
| 3 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0.25 | 0.2 |
| 5 | 0.05 | 0 | 0 | 0 | 0.45 |

表 5: 予測誤差が二番目に大きいケース（最大予測誤差 = 3.39）

(a) 予測情報

| タイプ | 入力満足度 | 予測満足度 | 正解 | 予測誤差 |
|---|---|---|---|---|
| 予測誤差 (大) | 5.0 | 4.39 | 1.0 | 3.39 |
| 予測誤差 (小) | 3.0 | 3.79 | 4.0 | 0.211 |

(b) 混同行列の全要素

| $\hat{y}\backslash y$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0.00135 | 0 | 0 | 0.00135 | 0.00270 |
| 2 | 0 | 0.00135 | 0 | 0.00270 | 0 |
| 3 | 0 | 0 | 0.00270 | 0.03374 | 0.00540 |
| 4 | 0 | 0 | 0.00810 | 0.44130 | 0.16464 |
| 5 | 0 | 0 | 0.00270 | 0.26451 | 0.06748 |

2：https://ollama.com/library/gemma3:27b

効に働く傾向を有するということを示唆する．

また，本提案手法の RMSE Loss は平均して 0.5608 であり，MAE Loss は平均して 0.4053 であった．一方で，ベースライン手法の RMSE Loss は平均して 0.7030 であり，MAE Loss は平均して 0.5191 であった．そしていずれの $p$ 値も 0.05 を下回ることから，本提案手法がベースラインと比較して予測誤差が有意に小さいことがわかる．また，我々は外れ値に強い MAE Loss が RMSE Loss よりも小さいことから，例外的なレビューケースの存在が精度に影響を及ぼしていると考えた．例外的なレビューケースと精度の関係を考察するため，最も予測誤差が大きいケースと次点で予測誤差が大きいケースを抽出する．また，最も予測誤差が小さいケースと次点で予測誤差が小さいケースも抽出する．

それぞれのケースを表 4，表 5，表 6 と表 7 に示す．表 4 と表 5 では，いずれのレビュアも混同行列に例外的なレビューケースが複数存在する．またいずれのレビュアにおいても，予測誤差が大きいケースは例外的なレビューケースを対象としたときである．これは，例外的なレビューケースに対する個人差の補正が本提案手法において有効ではないことを示唆している．しかし，例外的なレビューケースに該当しない満足度に対する個人差の補正ではいずれの予測誤差も 0.25 を切っている．この結果は，混同行列が構造的な個人差を捉える際に有効であることを示している．したがって，以上の結果は例外的なレビューケースのような大きな個人差が内在する主観評価を行うレビュアほど個人差の補正が困難であることを示唆している．また，表 6 と表 7 では，いずれのレビュアも混同行列に例外的なレビューケースが存在していない．そしていずれのレビュアにおいても，予測誤差が小さいケースはほぼ完全な個人差の正規化がなされている．これは，多数のレビュアが持つ平均的な判断基準に基づいた評価傾向を持つレビュアほど完全に近い形で個人差を補正することが可能であることを示唆している．これらの考察は，予測誤差の大きさに基づいて抽出した 4 種類のレビュアサンプルから得られたものである．そのため，得られた考察は抽出したサンプルに依存している可能性がある．しかしながら，例外的なレビューケースの割合や評価傾向における固有構造の有無によって補正効果が変化するという傾向は，混同行列による個人差補正の適用条件を示す重要な示唆であると考えられる．

表 6: 予測誤差が最小のケース（最小予測誤差 = 0.0003）

(a) 予測情報

| タイプ | 入力満足度 | 予測満足度 | 正解 | 予測誤差 |
|---|---|---|---|---|
| 予測誤差 (大) | 3.0 | 3.19 | 2.0 | 1.19 |
| 予測誤差 (小) | 5.0 | 4.0 | 4.0 | 0.0003 |

(b) 混同行列の全要素

| $\hat{y}\backslash y$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0.033 | 0 | 0 |
| 3 | 0 | 0 | 0.033 | 0.133 | 0 |
| 4 | 0 | 0 | 0.033 | 0.5 | 0.2 |
| 5 | 0 | 0 | 0 | 0.067 | 0 |

表 7: 予測誤差が二番目に小さいケース（最小予測誤差 = 0.0004）

(a) 予測情報

| タイプ | 入力満足度 | 予測満足度 | 正解 | 予測誤差 |
|---|---|---|---|---|
| 予測誤差 (大) | 3.0 | 3.16 | 2.0 | 1.16 |
| 予測誤差 (小) | 5.0 | 4.0 | 4.0 | 0.0004 |

(b) 混同行列の全要素

| $\hat{y}\backslash y$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0.01 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0.03 | 0 |
| 4 | 0 | 0 | 0.01 | 0.31 | 0.58 |
| 5 | 0 | 0 | 0 | 0.01 | 0.05 |

本実験では，主に二つの考察が得られた．一つ目は，本提案手法は主観評価に内在する個人差の解消というタスクにおいて有効に働く傾向があるということである．二つ目は，本提案手法は例外的なレビューケースのような大きな個人差の解消において改善する余地があるということである．したがって，提案手法が主観評価に内在する個人差の解消に有効であるという仮説は正しいと言える．

## 5 おわりに

本研究の目的は，リッカート尺度に基づく主観評価に内在する意味解釈の個人差を解消することである．我々は本研究の目的を達成するために，リッカート尺度に基づく個人差の内在した主観評価値を多数のレビュアに共通する個人差の内在しない参照尺度に基づく評価値へと正規化する手法を提案した．つまり本研究は，満足度予測タスクを予測ではなく主観評価に内在する意味解釈の個人差解消のための正規化タスクとして問題設定を再定義した．また本研究は，既存研究でノイズとして扱われる例外的なレビューケースの存在をレビュアの混同行列が保有する個人差として構造的に扱った．

我々は，提案手法の有効性確認のための実験を実施した．本実験では，ニューラルネットワークによる評価値の正規化タスクを用いて提案手法が主観評価に内在する個人差の解消に有効であるか検証した．実験結果より，本提案手法が主観評価に内在する個人差の解消というタスクにおいて有効に働く傾向にあることを確認した．一方で，例外的なレビューケースに対する個人差の解消は改善の余地があることを確認した．これらの結果より，意味解釈の個人差を評価傾向を表す混同行列の構造として扱うことによって，既存研究ではノイズとして扱われていた個人差が体系的に補正可能であることが示された．また，本提案手法の補正により主観尺度は共通した尺度上の客観的評価値として扱うことが可能となるため，評価値間の直接的な比較や

集計から導かれる分析結果の安定した解釈を可能とすることが示された．一方で本稿での実験結果より，本提案手法にはいくつかの懸念事項が存在している．

一つ目は，本提案手法の適用範囲が限定的なことである．本提案手法は，評価傾向を混同行列 $C^{(u_i)}$ によって表現した．また本提案手法は，事前学習済み LLM の Zero-Shot 推論による満足度予測結果 $\hat{y_j}$ を近似参照尺度として定義した．そのため，混同行列 $C^{(u_i)}$ や Zero-Shot 推論による満足度予測結果 $\hat{y_j}$ に関わる満足度予測の入力であるレビュー $r_j$ が手法全体の基盤となる．したがって，本提案手法はレビュー $r_j$ の影響を大きく受けてしまう．そのため我々は，本提案手法を用いる上で二つの条件が満たされる必要があったと考える．一つ目の条件は，レビュー $r_j$ が評価値 $y_j$ と対になっていることである．我々は EC サイトの満足度評価値において，評価理由を示すレビュー $r_j$ が存在せず評価値 $y_j$ が単体で投稿されているケースを本提案手法の適用範囲外とした．二つ目の条件は，レビュー $r_j$ が平均的な判断傾向のもとで解釈可能であるということである．レビュー $r_j$ が平均的な判断傾向のもとで解釈不可能である場合，レビュー $r_j$ を基に付与された満足度 $y_j$ の意味解釈における違いを個人差という範囲で扱うことは難しい．そのため我々は，平均的な判断枠組みのもとで解釈することが困難なレビューや満足度評価を実施しているレビュアは本提案手法の適用範囲外とした．具体的には，平均的に些細とされる程度の欠点を過剰に批判するレビュアや，個人的な信念や考えに強く依存するレビューを書くレビュアを指す．また，我々は表 2c で示すようなレビュー数が極端に少ないレビュアは評価傾向の解釈が不安定になるため，本提案手法の適用範囲外とした．以上のように限定された適用範囲を拡張することによって，より実用的な問題設定を実施可能になると考える．

二つ目は本稿において我々が，近似参照尺度を LLM による Zero-Shot 推論の予測値で定義したことである．本稿において我々は，追加学習や参考情報の添付を実施していない事前学習済み LLM が Zero-Shot 推論によって妥当な評価を返すという仮定を下に参照尺度を定義した．しかし，我々はこの仮定の妥当性を確認できていない．そのため，我々はこの仮定の妥当性を確認するか，近似参照尺度を LLM による Zero-Shot 推論の予測値以外で定義する必要があると考える．我々は前者に関して，人力のアノテーションとそれに伴う多数決を利用することによって，仮定の検証が可能であると考える．また我々は後者に関して，複数の LLM を用いた Zero-Shot 推論の多数決や不特定多数のレビュアが書いたレビューをランダムに参考データとした Few-Shot Prompting によって解決可能だと考える．

三つ目は，例外的なレビューケースに対する個人差の解消方法である．本稿の実験において，例外的なレビューケースに対する個人差の解消には改善の余地があることを確認した．我々は，例外的なレビューケース要素に対する重みづけを改善アプローチとして提案する．例外的なレビューケースは，基本的に他要素と離れた位置に存在している．そして，レビュアが持つ例外的なレビューケース割合は平均して 5.93%程度である．これらのことより，我々は混同行列内の例外的なレビューケース要素がモデル推論時に重要視されていないと考える．したがって，我々は例外的なレビューケース要素に対する重みづけが有効なアプローチとなるのではないかと考えた．

四つ目は，本提案手法の汎化的な適用範囲である．本稿では，5 段階のリッカート尺度を対象として様々な実験を実施した．そのため，7 段階のリッカート尺度や満足/不満足のみの 2 段階尺度に対しては，本稿の実験結果や考察が当てはまるかが定かではない．特に例外的なレビューケースの定義という点においては，本提案手法を拡張する必要がある．したがって，我々は尺度段階が増加した場合においての例外的なレビューケースの定義や前述した例外的なレビューケース要素に対する適切な重みづけを改めて調査・分析する必要があると考える．

## 謝　　辞

本研究の一部は JSPS 科研費（24K03044，23K28383）の助成を受けたものです．本研究では，NII IDR データセット提供サービスにより楽天グループ株式会社から提供を受けた「楽天データセット」（https://rit.rakuten.com/data_release/）を利用しました．

## 文　　献

[1] Sabahattin Yeşilçınar and Mehmet Şata. Examining rater biases of peer assessors in different assessment environments. *International Journal of Psychology and Educational Studies*, Vol. 8, No. 4, pp. 136–151, 2021.

[2] Harry Helson. Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychological review*, Vol. 55, No. 6, p. 297, 1948.

[3] Edward L Thorndike, et al. A constant error in psychological ratings. *Journal of applied psychology*, Vol. 4, No. 1, pp. 25–29, 1920.

[4] Kushal Anjaria. Knowledge derivation from likert scale using z-numbers. *Information Sciences*, Vol. 590, pp. 234–252, 2022.

[5] Lotfi A Zadeh. A note on z-numbers. *Information sciences*, Vol. 181, No. 14, pp. 2923–2932, 2011.

[6] Bingkun Wang, Shufeng Xiong, Yongfeng Huang, and Xing Li. Review rating prediction based on user context and product context. *Applied Sciences*, Vol. 8, No. 10, p. 1849, 2018.

[7] 坂本新真, 牛尼剛聡. 評価値予測タスクを用いたユーザの嗜好に適合したレビューの抽出. 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM2023), pp. 1–8. 一般社団法人電気情報通信学会, 2023.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.