授業評価に対するイロレーティングの適用とその予測の試み

加納 豊之　角田 雅照

大学において，学生による授業評価は広く実施されている．ただし，学生によっては全ての評価項目に対し，5 段階ならば 3（普通）をつけるといった評価を行う可能性がある．逆に，授業に対して強い不満を持つ学生は，全ての評価項目に対し機械的に最低点をつける可能性がある．本研究では，このような授業評価の偏りを低減させるために，一対比較法を繰り返して行い，その結果からイロレーティングを算出することを試みる．実験ではイロレーティングによる授業ランキングの納得度を被験者に評価してもらった．その結果，イロレーティングと 10 段階評価の平均順位を用いたランキングの評価が最も高かった．さらに，協調フィルタリングを用いて，一対比較法による授業評価がどの程度の精度で予測可能であるかも確かめた．

In the university, student evaluation of teaching (SET) is widely applied. However, some students may evaluate all items (e.g., progress speed of lecture) as medium on SET. Also, when students do not satisfy the teaching at all, they may evaluate all items as worst. To avoid such evaluation bias, we apply pairwise comparison and Elo ratings, and make teaching ranking. In the experiment, subjects evaluated the teaching ranking based on Elo ratings. As a result, they highly evaluated the ranking based on average rank of Elo ratings and ten-scale evaluation. Also, we used collaborative filtering to predict the pairwise comparisons, and evaluated the prediction accuracy.

1　はじめに

大学において，学生による授業評価は広く実施されている．典型的な授業評価では，「講義時の教員の説明の速さは適切か」などの複数の評価項目に対し，10 段階の数値などにより評価する．また，総合評価についても同様に 10 段階の数値などで得点がつけられる．学生による授業評価により，教員は講義において改善すべき点などを客観的に知ることができる．これにより講義が改善され，教育内容の質が高まることが期待される．

ただし，場合によっては授業評価が正常に機能していない可能性がある．例えば，アンケートが億劫だと感じた学生は，全ての評価項目に対し，5 段階ならば 3（普通）をつけるといった評価を行う可能性がある．逆に，授業に対して強い不満を持つ学生は，全ての評価項目に対し機械的に最低点をつける可能性がある．このような極端な例でなくても，授業評価に関して絶対的な基準が定義されていないことが多いため，評価は必ずしも簡単ではない．

本研究の最終的なゴールは，このような授業評価の偏りや評価の困難さを低減させることにより，評価の信頼性を高め，講義内容の改善を推し進めることである．そのために本研究では一対比較法に着目し，授業評価に用いる．一対比較法では，2 つの評価対象のうち，どちらが良いかにより評価を行う．一対比較法を授業評価に取り入れている研究は存在するが（文献[4] など），本研究は一対比較法を繰り返して行い，その結果からイロレーティングを算出することを試みる．イロレーティングはチェスなどの一対一のゲームのプレイヤーの強さを表す指標であり，過去の対戦成績に基づいて計算される．イロレーティングは強い相手に勝った場合を重視し，逆に弱い相手に勝った場合は重視しないという特徴がある．このイロレーティン

Application of Elo Ratings and Its Prediction to Student Evaluation of Teaching
Toyoyuki Kano, Masateru Tsunoda, 近畿大学理工学部, Department of Informatics, Kindai University.

グを用いることにより，授業評価の信頼性が高まることが期待される．

実験ではイロレーティングにより授業をランキングし，その後，そのランキングの納得度を被験者に評価してもらった．さらに，実験では協調フィルタリングを用いて，一対比較法による授業評価がどの程度の精度で予測可能であるかも確かめる．これは，授業評価のデータが欠けている場合に，どの程度補完が可能なのかを明確にするためである．

授業をランキングする目的は，現在の相対的な評価を知ることにより，授業改善のきっかけとするためである．評価が標準的であっても，授業改善の余地がないとはいえない．例えば著者らの所属校では，10 段階評価で 7 点を標準として学生に授業を評価してもらっているが，平均点が 7 点の授業は，おおむね全授業の下位 25%程度に該当することが多い．実際に著者のひとりが担当した授業 2 つは，初年度は 7 点であったが全授業の下位に相当した．それをきっかけとして授業の改善（演習問題を多数実施し，授業の要点を明確にするなど）を行うことにより，翌年度には平均点が 0.5 点程度改善し，全授業の中位程度となった．

以降，2 章では授業ランキングについて説明し，3 章では協調フィルタリングについて述べる．4 章では実験の方法について説明し，5 章では実験結果について述べる．最後に 6 章でまとめを述べる．

2 授業ランキング

2.1 一対比較法

学生による授業評価は，授業の最終回に 5 段階や 10 段階などの得点を用いて行われることが多い．例えば，1 を最低点，10 を最高点，7 を標準点とし，学生が授業を評価する．ただし，学生が十分に考慮せずに標準点で授業を評価したり，1 点や 2 点の極端に低い点数で授業を評価したりする可能性がある．

そこで本研究では，一対比較法に着目する．一対比較法とは，ある 2 つの対象について，どちらが好ましいかにより評価する方法である．表 1 に一対比較法による評価例を示す．例えば，授業 A と B はどちらが優れているか，授業 A と C ではどちらか，授業 B と C では，というように評価を行う．ランダムにアイテム（画像）をピックアップし，一対比較法を繰り返して多数の評価データを集めることにより，多くのユーザが好ましいと考えるアイテムを特定する方法が提案されている [3]．本研究では同様の方法により，授業評価を試みる．

表 1　10 段階評価と一対比較法による評価の例

(a) 10 段階評価

授業	評価
A	7
B	8
C	6
D	9
・・・	・・・

(b) 一対比較法による評価

授業 1	授業 2	好ましい授業
A	B	A
A	C	A
B	C	C
C	D	C
・・・	・・・	・・・

2.2 イロレーティング

一対比較法の結果を単純に順位付けに用いると，優劣の結果を集計することになる．例えば，授業 A は 10 回中 8 回「優れている」との評価を受けた，授業 B は 10 回中 5 回同様の評価を受けた，などと集計する．ただし，一対比較法によるランキングでは，全ての組み合わせについて評価することは難しい．そのため，授業 A は元々評価の低い授業としか比べられておらず，逆に授業 B は評価がトップランクの他授業と比べられている可能性もあり，単純な集計では授業を正しくランキングできない可能性がある．

そこで本研究ではこの問題を解決するため，文献 [3] と同様にイロレーティング [1] を用いる．ある対象を主観により評価するという観点においては文献 [3] と同様であるため，イロレーティングが有効に働く可能性がある．イロレーティングはチェスなどの一対一の対戦ゲームの勝敗を元に，プレイヤーの実力を順位付けする方法である．イロレーティングでは，実力に大きな差がある場合の勝利と，差がない場合について，同等に扱わない．例えば，ランキング上位のプレイヤーに勝利した場合，その勝利を重視し，逆にランキング下位のプレイヤーに対する勝利は重視せず，各プレイヤーのランキングが順次更新される．ランキング上位か下位かは過去の対戦結果（レーティングの結果）に基づく．イロレーティングではレーティング（スコア）が算出され，レーティングが大きいほど

優れていることを示し，レーティング順に順位づけされる．

2.3 関連研究

一対比較法を授業評価に取り入れている研究は，文献[4]など，いくつか存在する．ただし，これらの研究では，階層分析法（AHP; Analytic Hierarchy Process）において一対比較法を適用しており，一対比較法に基づいてイロレーティングを計算するなどは行っていない．

イロレーティングを教育分野に適用した研究は数多く存在する（文献[5]など）．ただし，これらの研究では学生の理解度を評価するために用いており，授業評価には用いていない．

2.4 一対比較法に基づくランキング

2.4. 本研究では以下の 3 種類の方法により，一対比較法に基づく授業の順位づけを試みる．

- イロレーティングに基づくランキング: イロレーティングのレーティングに基づき，各授業を順位付けする．
- 平均順位に基づくランキング: イロレーティングによる授業の順位と，10 段階などの点数による授業の順位の平均値を算出し，それを順位とする．例えば授業 A において，イロレーティングによる順位が 2 位，点数による順位が 4 位とすると，$(2+4)\ /\ 2 = 3$ 位を順位とする．
- 平均評価に基づくランキング: ある授業に対する，イロレーティングによるレーティングと，10 段階などの点数の平均値を算出し，その値に基づき順位づけする．平均値の計算前に，イロレーティングによるレーティングと 10 段階などの点数は，以下の式を用いて値域が $[0, 1]$ にされる．

$$r' = \frac{r - \min(R)}{\max(R) - \min(R)} \qquad (1)$$

ここで，r は変換対象の値，$\max(R)$ と $\min(R)$ はそれぞれ R を含む変数の最大値，最小値，r' は変換後の値を表す．この計算方法は，値域を変換する際に広く用いられる方法の 1 つである[7]．

3 協調フィルタリングによる予測

協調フィルタリングは，ユーザにとって好ましい，または役立つと考えられるアイテム（書籍，音楽など）を推薦するための手法として用いられている[2][6][8]．「協調」とは，ユーザの知識を利用することを意味し，「フィルタリング」とは，大量のアイテムの中から，役立つアイテムだけを選び出して推薦することを意味する．一般的な協調フィルタリングで推薦を行う場合，各ユーザが各アイテムを（5 段階の数値などで）評価していることが前提となる（システムによっては，ユーザがそのアイテムを閲覧したかどうかを評価の代わりに用いることもある）．あるユーザが未評価のアイテムが，そのユーザにとって役立つと考えられる場合，そのアイテムを推薦する．

協調フィルタリングの主なアルゴリズムとして，ユーザベース手法とアイテムベース手法の 2 つがある．ユーザベース手法は，「アイテムの評価（好み）が似たユーザは，どのアイテムに対しても似た評価を行う」と仮定し，推薦を行う．具体的には，各ユーザの各アイテムに対する評価を要素とするベクトルを，ユーザごとに作成し，そのベクトルのなす角をユーザの類似度とする．そして推薦対象のユーザが未評価で，かつ類似したユーザの評価が高いアイテムを推薦する．ユーザベース手法を用いた推薦システムとして，Resnick ら[8] の GroupLens が挙げられる．GroupLens は，Usenet にある多数のニュース記事から，ユーザの好みに合うと予測される記事を選び出して推薦するシステムである．

アイテムベース手法は Sarwar ら[6] によって提案されたアルゴリズムであり，アイテム間の類似度に基づいて推薦を行う．アイテムベース手法の場合も，各ユーザの各アイテムに対する評価を要素としてベクトルを作成するが，ユーザごとにベクトルを作成するのではなく，アイテムごとに作成し，類似度を計算する．すなわち，「あるグループのユーザに高評価されるアイテムは，類似の性質を持っている」と仮定し，推薦対象のユーザが高い評価を行っているアイテムと類似度の高いアイテムを推薦する．

本研究では，協調フィルタリングを用いて授業評価

表 2　協調フィルタリングによる予測時に想定するデータ

	t_1	t_2	…	t_j	…	t_n
u_1	r_{11}	r_{12}	…	r_{1j}	…	r_{1n}
u_2	r_{21}	r_{22}	…	r_{2j}	…	r_{2n}
…	…	…	…	…	…	…
u_i	r_{i1}	r_{i2}	…	r_{ij}	…	r_{in}
…	…	…	…	…	…	…
u_m	r_{m1}	r_{m2}	…	r_{mj}	…	r_{mn}

の一対比較結果の予測を行う．予測では，表 2 のような $m \times n$ のマトリックス形式のデータを想定する．$u_i \in \{u_1, u_2, ..., u_m\}$ は i 番目の評価者（学生）を表し，$t_j \in \{t_1, t_2, ..., t_n\}$ は j 番目の授業科目の一対比較結果を表す．例えば，t_1 は授業 A と B を一対比較した場合，t_2 は授業 A と C を一対比較した場合を表す．また，$r_{ij} \in \{r_{11}, r_{12}, ..., r_{mn}\}$ は授業科目の一対比較結果を表す．例えば，授業 A と B を一対比較し，A のほうが優れていると評価した場合は r_{ij} を 1，B と評価した場合は 0 とする．

4　実験

4.1　概要

実験では，授業の評価データを収集し，そのデータに基づき授業のランキングを作成し，被験者にランキングの納得度を評価してもらった．また，協調フィルタリングを用いて，授業の評価の予測を試みた．

実験の目的を明確にするために，本研究では以下の 2 つのリサーチクエスチョンを設定した．

- RQ1: 一対比較法に基づく授業のランキングと，従来法に基づくランキングでは，どちらの納得度が高いのか?
- RQ2: 一対比較法による評価結果は，どの程度の精度で予測できるのか?

本研究では，納得度は「授業評価に基づくランキング結果として，納得できるかどうか」と定義した．次節で述べるように，被験者には実際に所属校で実施しているアンケートと同様の基準で回答するよう指示しており，納得度の高いランキングのほうが授業改善に有用であるといえる．

一部の学生が授業評価において回答しないことがあるため，全てのデータが収集できない場合がある．この場合，予測によりどの程度補完が可能なのかを確かめるために RQ2 を設定した．

4.2　授業評価データ

授業の評価データを収集するために，学部 3- 4 年生の被験者 16 人からデータを収集した．一対比較法による評価，10 段階評価（最低点は 1，最高点は 10，標準点は 7）のどちらについても，ランダムに科目を表示し，それぞれ 20 件以上（一対比較法の場合は 20 回以上）の授業を評価してもらった．一対比較法による評価，10 段階評価の順番は被験者により入れ替えを行った．

10 段階評価は著者らの所属校でも同様に実施しているものであり，被験者に対しても，通常の授業アンケートと同様の基準で答えるように指示した．このアンケートでは，評価に際して技術習得度などの特定の観点を指定しておらず，総合的な満足度により評価をしている．このため，何を重視して総合的な評価をするかは個人差があると考えられるが，同じ被験者に対して 10 段階評価と一対比較法による評価をしてもらっているため，個人差の影響はある程度吸収されていると考えられる．なお，この個人差が大きいほど，納得度も低下すると思われるため，納得度の高いランキングは，この個人差をより吸収できている可能性がある．

評価対象の授業は 54 件とした．データ収集の結果，回答数が 3 件以上となった授業は 32 件であった．このデータを以降の実験では用いる．

4.3　授業ランキング

RQ1 に答えるために，一対比較法に基づく授業のランキングと，従来の 10 段階評価に基づく授業のランキングを作成し，被験者にどのランキングの納得度が高いかを答えてもらった．ランキングは以下の 4 種類を作成した．

1. 10 段階評価に基づくランキング
2. イロレーティングに基づくランキング
3. 平均順位に基づくランキング

4. 平均評価に基づくランキング

1. は従来法によるランキング，2. から 4. は 2.4 節で説明した方法に基づくランキングである．

被験者には，これらのランキングの納得度についても一対比較法により評価してもらった．すなわち，一対比較法の結果から，イロレーティングによりレーティングを算出し，このレーティングが高いランキングを納得度が高いものとした．

4.4 評価の予測

RQ2 に答えるために，一対比較法による授業評価を，協調フィルタリングを用いて，3 章で説明した方法により予測した．予測時にはユーザベースとアイテムベースのそれぞれを用い，精度を確かめた．具体的には，以下の 4 種類の予測を行った．

- 一対比較評価の予測（ユーザベース）
- 一対比較評価の予測（アイテムベース）
- 10 段階評価の予測（ユーザベース）
- 10 段階評価の予測（アイテムベース）

ベンチマークとするために，従来の 10 段階評価についても協調フィルタリングにより予測した．10 段階評価は協調フィルタリングでしばしば用いられるデータと同様であり，予測精度が比較的高いことが期待されるが，一対比較法による授業評価データは欠損値が多くなるため，予測精度が低くなる可能性がある．

予測時にはリーブワンアウト法を用い，予測値と実測値の絶対誤差を算出して予測精度を評価した．例えば，評価の予測値が 8 で実測値が 6 の場合，絶対誤差は 2 となる．また，予測値に基づき授業ランキングも作成し，実際のランキングの順位との絶対誤差を評価した．例えば予測値に基づく授業 A の順位が 5 位，実測値に基づく授業 A の順位が 3 位の場合，絶対誤差は 2 となる．

5 結果

5.1 授業ランキングの比較

被験者による，4 種類の授業ランキングの納得度の評価を表 3 に示す．表において，「勝」は一対比較において他方より優れていると評価された回数，「負」は逆に劣っていると評価された回数である．

表 3 各ランキングに対する評価

ランキング法	レーティング	比較回数	勝	負
平均順位	2226	40	25	15
10 段階	2209	40	16	24
イロレーティング	2206	40	21	19
平均評価	2159	40	18	22

表 4 評価予測時の誤差

予測対象(予測方法)	絶対誤差平均	絶対誤差中央値
一対比較（ユーザベース）	0.39	0.0
一対比較（アイテムベース）	0.43	0.0
10 段階（ユーザベース）	1.35	1.0
10 段階（アイテムベース）	1.47	1.0

表 5 予測に基づくランキングの誤差

予測対象(予測方法)	絶対誤差平均	絶対誤差中央値
一対比較（ユーザベース）	5.06	3.5
一対比較（アイテムベース）	5.50	3.0
10 段階（ユーザベース）	3.88	2.5
10 段階（アイテムベース）	3.66	2.0

各ランキングのレーティングに大きな差はなかったが，平均順位に基づくランキングの評価が最も高く，平均評価に基づくランキングの評価が最も低かった．10 段階評価に基づくランキングはイロレーティングに基づくランキングより評価が高かったが，レーティングの差はごくわずかであった．

これらの結果より，10 段階評価とイロレーティングの平均順位を用いて授業をランキングすることにより，もっとも納得度が高くなる可能性がある．すなわち，RQ1 に対する答えは，「両者の順位を平均したランキングが最も納得度が高い」となる．ただし，被験者は 16 人であり，一対比較のデータも充分に収集されているとまではいえないため，結果の信頼性を高めるためには，更なるデータ収集が必要である．

5.2 予測精度の比較

協調フィルタリングを用いて，一対比較法と 10 段階による授業評価を予測した場合の精度を表 4 に示す．一対比較法による評価の予測においてはユーザ

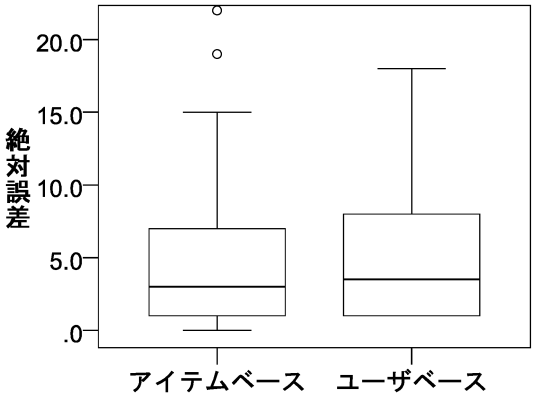

図 1　一対比較評価の予測に基づくランキングの誤差

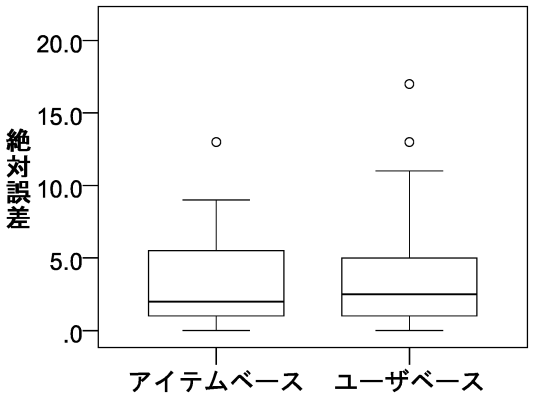

図 2　10 段階評価の予測によるランキングの誤差

ベースのほうが比較的精度が高く，10 段階の場合についても同様であった．絶対誤差は一対比較法のほうが小さいが，10 段階の場合は値域は $[1, 10]$，一対比較法の値域は $[0, 1]$（ある授業が好ましいかどうかが 0 と 1 で表される．3 章参照）であることから，後者の精度はあまり高くないといえる．

授業評価の予測値に基づいてランキングを作成した場合の順位の誤差を表 5 に示す．また，それぞれの誤差の箱ひげ図を図 1，図 2 に示す．この場合，一対比較法に関して，誤差の中央値に着目するとアイテムベースのほうが精度が高く，10 段階に関しても同様にアイテムベースの精度が高かった．箱ひげ図でも同様に，アイテムベースのほうが若干データのばらつきが小さかった．

表 5 において，一対比較法と 10 段階のランキングの誤差を比較すると，10 段階のほうが小さかった．ただし，一対比較法と 10 段階の誤差の違いは 1.5 程度であり，差は無視できないが非常に大きいとまではいえない．ランキング対象の授業が 32 件で，一対比較法の誤差平均が 5 程度であったことから，RQ2 に対する答えは「評価結果の予測は不可能ではないが，充分実用的な精度であるとまではいえない」となる．

6　おわりに

本研究では，学生による授業評価において，一対比較法による評価を行い，その後イロレーティングにより授業をランキングすることを試みた．イロレーティングによるランキングでは，3 種類の方法（イロレーティングをそのまま用いる，10 段階評価との平均順位を用いる，10 段階評価との平均評価を用いる）でランキングを作成した．その結果，10 段階評価との平均順位を用いたランキングの納得度が最も高くなった．

また，協調フィルタリングにより，一対比較法による評価を予測し，その結果に基づいてランキングを作成し，どの程度の誤差となるのかを確かめた．その結果，一対比較法による評価は予測可能であるが，精度に関しては改善の余地があることがわかった．

本研究の主要な貢献は，学生による授業評価に対してイロレーティングを適用することにより，授業評価の結果をより適切に反映した授業ランキングを作成できる可能性を示唆したことである．なお，PBL(Project Based Learning) 型講義の評価への適用時には，学生の満足度という観点から評価するならば，必ずしも PBL 形式の講義同士を比較する必要はなく，実習や通常の座学形式の講義と比較しても構わない．

今後の予定は，更に被験者を増やすとともに評価データを多く収集し，実験結果の信頼性を高めることである．また，授業により異なる学習目標とそれを理解した学生かどうかを考慮した分析も必要であると考える．

謝辞 本研究の一部は，文部科学省科学研究補助費（基盤 C：課題番号 16K00113）による助成を受けた．

参 考 文 献

[1] A. Elo: The Rating of Chess Players, Past and Present, Ishi Press, 2008.

[2] D. Goldberg, D. Nichols, B. Oki, and D. Terry: Using Collaborative Filtering to Weave an Information Tapestry, Communications of the ACM, vol.35, no.12 (1992), pp.61-70.

[3] S. Hacker, and L. Ahn: Matchin: eliciting user preferences with an online game, In Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI), 2009, pp.1207-1216.

[4] 池原一哉，豊田秀樹: 評価基準の重要度評定と学生による授業の一対比較評定を統合する授業評価モデルの提案:一学生による評価と教員による評価の比較・検討一，教育心理学研究，vol.60，no.1 (2012)，pp.48-59.

[5] R. Pelnek, Applications of the Elo rating system in adaptive educational systems, Computers & Education, vol.98 (2016), pp.169-179.

[6] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl: Item-Based Collaborative Filtering Recommendation Algorithms, Proc. of International World Wide Web Conference (WWW10), 2001, pp.285-295.

[7] K. Strike, K. El Eman, and N. Madhavji: Software Cost Estimation with Incomplete Data, IEEE Transactions on Software Engineering, vol.27, no.10 (2001), pp.890-908.

[8] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, Proc. of Conference on Computer Supported Cooperative Work (CSCW), 1994, pp.175-186.