# Calibrating the Predictions for Top-N Recommendations

| | |
|---|---|
| 著者 | Masahiro Sato (Fuji Film) |
| 発行年 | 2024 |
| 学会名 | RecSys |
| キーワード | Calibration Ranking Recommend Top-N |
| URL | https://arxiv.org/abs/2408.11596 |

## 概要

- ユーザの嗜好の予測における予測値は較正させること（キャリブレーション）が必要である
- たとえば，推薦システムが予測値を0.5と出力した場合，そのような予測を持つ10品目のうち，5品目が買われていれば，適切にキャリブレーションされたと言える
- この研究では上位N個の推薦アイテムにおけるミスキャリブレーションに対処することを目的とする
- この目的に対する評価手法を定義し，上位N項目に着目したキャリブレーションモデルの最適化手法を提案する

## 研究背景

- 推薦モデルはユーザの過去の評価や行動に基づいて，ユーザが興味を持ちそうなアイテムを予測し，上位N件を提示するシステム
- キャリブレーションとは予測スコアを信頼できる値に較正することであり，現実との整合性が取れていると，よくキャリブレーションできたと言える
- 実際に使われるのは上位N件であることがほとんどだが，既存研究では全てのアイテムに対するキャリブレーション精度しか見ていない

## 論文の肝

- Top-Nアイテムに対するキャリブレーションを評価する手法の提案
    - これにより既存手法が全アイテムでは良い性能を出しているが，Top-Nに対してはミスキャリブレーションしていることがわかった
- Top-Nアイテムに着目したキャリブレーション手法を提案
- 従来手法に比べ安定して高い性能を実証することができた

# 前提知識

キャリブレーション誤差 (ECE: Expected Calibration Error)

- $y_{ui}$: ユーザuとアイテムiに対するフィードバック（評価点やクリックなどの行動）
- $s_{ui}$: レコメンドが出力するフィードバックスコア
- $\hat{y}_{ui} = g(s_{ui})$: キャリブレーションモデルに補正された予測値

このような定義に対してキャリブレーション誤差CEは

$$CE = |P(y = 1|\hat{y}) - \hat{y}|$$

で表される．実際にはCEはフィードバックの期待値で計算することができ，

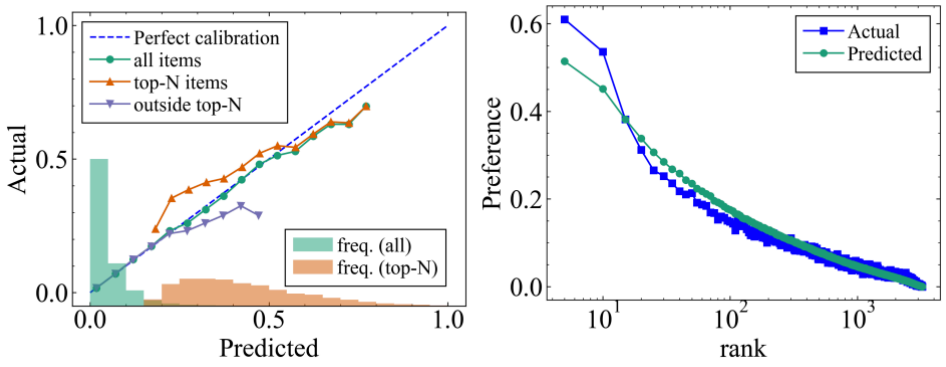
$$CE = |E[y|\hat{y}] - \hat{y}|$$

CEの期待値ECEはサンプルをビニングすることで

$$ECE = E[CE] \approx \sum_{m=1}^{M} \frac{|B_m|}{n} |\frac{\sum_{k \in B_m} y_k}{|B_m|} - \frac{\sum_{k \in B_m} \hat{y_k}}{|B_m|}|$$

によって計算される．

これまでの研究では全アイテムに対してキャリブレーション誤差を計算していたため，全アイテムの平均では良い結果が得られていたが，上位に絞ってみると過小評価が目立っている




(a) Reliability diagram (N=20) (b) Rank-based calibration plot

**Figure 1: Calibration plots for NCF with Gaussian calibration applied to preference prediction in the Kuairec dataset.**

# 提案手法

## 評価指標

- ECE@N: ECEのうちN個だけで評価したもの
- RDECE@N: 高ランクは重要度が高いことから

$$RDECE = \frac{N}{\sum_r w_r} \sum_{r=1}^{N} \frac{w_r |B'_r|}{n'} \left| \frac{\sum_{k \in B'_r} y_k}{|B'_r|} - \frac{\sum_{k \in B'_r} \hat{y}_k}{|B'_r|} \right|, w_r = 1/r$$

としてランクが高い方が評価の重みが大きくなるように評価

## キャリブレーション手法

- Top-Nについて抽出
- ランクに基づいてグループ分け
- 上位ほど大きくなるような重みをづけた損失で学習

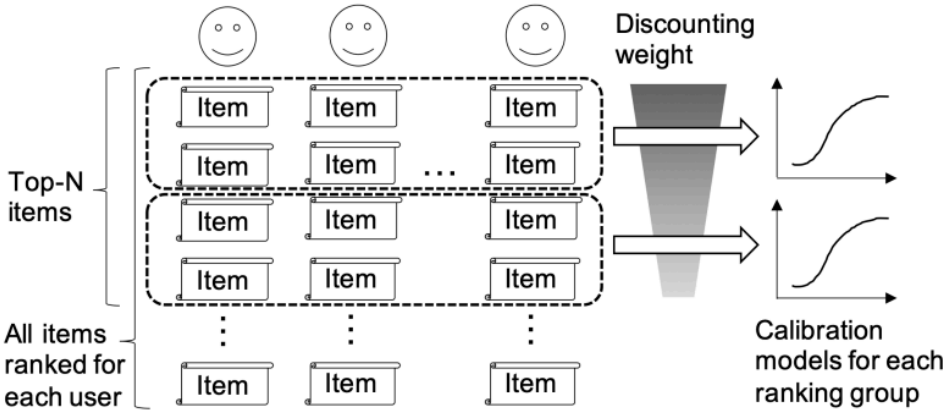



**Figure 2: Proposed calibration method. Top-N items are grouped by their ranks. Then calibration models for each ranking group are trained with weights that decrease with the ranks of each training sample.**

# 評価実験

- **データセット**
  - MovieLens-1M
  - KuaiRec Dataset
    - 1411人のユーザと3327のアイテム全てに接した完全観測データ
- **推薦システム**
  - itemKNN
  - 行列分解アルゴリズム
  - BPR
    - ベイズパーソナライズランキング
  - NCF
    - ニューラル協調フィルタリング
  - LightGCN
- **キャリブレーションモデル**
  - **非パラメトリック:** Histogram Binning, Isotonic Regression
  - **パラメトリック:** Platt Scaling, Beta Calibration, Gaussian/Gamma Calibration

- **Vanilla:** キャリブレーションなしの元のスコア

- **比較モデル**
  - 提案手法: **TNF** (Top-N Focus)
  - **VAD** (variance-adjusting debiasing)既存のTop-Nスコアの過大評価を補正する手法

- **結果**
  - TNFが全ての組み合わせで最も良い評価を達成している
  - VADはTop20で見た時に誤差を増幅させる傾向にある

**Table 1: Calibration comparison in the ML-1M dataset with itemKNN and MF. The best results among compared methods (Original, VAD, and TNF) for each combination of recommenders and calibration models are highlighted in bold.**

| | | ECE@20 | | | RDECE@20 | | |
|---|---|---|---|---|---|---|---|
| Recommender | Calibration model | Original | VAD | TNF | Original | VAD | TNF |
| itemKNN | Vanilla (as reference) | 0.075 ± 0.010 | 0.524 ± 0.013 | - | 0.069 ± 0.021 | 0.605 ± 0.037 | - |
| | Histogram binning | 0.194 ± 0.011 | 0.471 ± 0.010 | **0.045** ± 0.006 | 0.229 ± 0.045 | 0.537 ± 0.038 | **0.072** ± 0.022 |
| | Isotonic regression | 0.177 ± 0.011 | 0.462 ± 0.010 | **0.026** ± 0.015 | 0.197 ± 0.044 | 0.521 ± 0.038 | **0.070** ± 0.021 |
| MF | Vanilla (as reference) | 0.098 ± 0.014 | 0.327 ± 0.016 | - | 0.079 ± 0.020 | 0.332 ± 0.019 | - |
| | Histogram binning | 0.065 ± 0.013 | 0.351 ± 0.014 | **0.029** ± 0.010 | 0.089 ± 0.014 | 0.387 ± 0.014 | **0.027** ± 0.008 |
| | Isotonic regression | 0.036 ± 0.012 | 0.325 ± 0.013 | **0.017** ± 0.008 | 0.047 ± 0.010 | 0.348 ± 0.013 | **0.027** ± 0.009 |

**Table 2: Calibration comparison in the Kuairec dataset with NCF, lightGCN, and BPR. The best results among compared methods (Original, VAD, and TNF) for each combination of recommenders and calibration models are highlighted in bold.**

| | | ECE@20 | | | RDECE@20 | | |
|---|---|---|---|---|---|---|---|
| Recommender | Calibration model | Original | VAD | TNF | Original | VAD | TNF |
| NCF | Vanilla (as reference) | 0.340 ± 0.013 | 0.171 ± 0.016 | - | 0.292 ± 0.018 | 0.115 ± 0.021 | - |
| | Histogram binning | 0.260 ± 0.011 | 0.312 ± 0.011 | **0.020** ± 0.004 | 0.342 ± 0.014 | 0.394 ± 0.014 | **0.023** ± 0.005 |
| | Isotonic regression | 0.080 ± 0.017 | 0.152 ± 0.015 | **0.012** ± 0.006 | 0.116 ± 0.021 | 0.197 ± 0.021 | **0.023** ± 0.005 |
| | Platt scaling | 0.040 ± 0.008 | 0.096 ± 0.021 | **0.013** ± 0.003 | 0.070 ± 0.024 | 0.152 ± 0.028 | **0.023** ± 0.005 |
| | Beta calibration | 0.093 ± 0.018 | 0.147 ± 0.013 | **0.010** ± 0.004 | 0.102 ± 0.017 | 0.183 ± 0.018 | **0.023** ± 0.005 |
| | Gaussian calibration | 0.089 ± 0.024 | 0.139 ± 0.024 | **0.016** ± 0.005 | 0.093 ± 0.031 | 0.175 ± 0.032 | **0.023** ± 0.005 |
| | Gamma calibration | 0.087 ± 0.020 | 0.224 ± 0.038 | **0.013** ± 0.007 | 0.119 ± 0.029 | 0.279 ± 0.045 | **0.026** ± 0.006 |
| lightGCN | Vanilla (as reference) | 0.554 ± 0.008 | 0.446 ± 0.007 | - | 0.472 ± 0.007 | 0.363 ± 0.007 | - |
| | Histogram binning | 0.257 ± 0.009 | 0.297 ± 0.009 | **0.019** ± 0.004 | 0.339 ± 0.008 | 0.380 ± 0.008 | **0.022** ± 0.003 |
| | Isotonic regression | 0.070 ± 0.009 | 0.128 ± 0.008 | **0.014** ± 0.004 | 0.078 ± 0.010 | 0.152 ± 0.009 | **0.022** ± 0.004 |
| | Platt scaling | 0.132 ± 0.009 | 0.184 ± 0.008 | **0.018** ± 0.006 | 0.164 ± 0.009 | 0.227 ± 0.008 | **0.022** ± 0.004 |
| | Beta calibration | 0.097 ± 0.009 | 0.145 ± 0.008 | **0.015** ± 0.007 | 0.092 ± 0.010 | 0.165 ± 0.009 | **0.022** ± 0.004 |
| | Gaussian calibration | 0.096 ± 0.011 | 0.152 ± 0.012 | **0.016** ± 0.005 | 0.112 ± 0.013 | 0.182 ± 0.013 | **0.022** ± 0.004 |
| | Gamma calibration | 0.131 ± 0.011 | 0.185 ± 0.011 | **0.015** ± 0.005 | 0.164 ± 0.012 | 0.228 ± 0.012 | **0.022** ± 0.004 |
| BPR | Vanilla (as reference) | 0.446 ± 0.005 | 0.429 ± 0.005 | - | 0.361 ± 0.004 | 0.344 ± 0.004 | - |
| | Histogram binning | 0.291 ± 0.006 | 0.306 ± 0.006 | **0.018** ± 0.004 | 0.383 ± 0.005 | 0.399 ± 0.004 | **0.020** ± 0.003 |
| | Isotonic regression | 0.034 ± 0.006 | 0.046 ± 0.006 | **0.011** ± 0.005 | 0.064 ± 0.004 | 0.077 ± 0.004 | **0.019** ± 0.004 |
| | Platt scaling | 0.143 ± 0.006 | 0.153 ± 0.006 | **0.016** ± 0.004 | 0.186 ± 0.005 | 0.197 ± 0.005 | **0.020** ± 0.003 |
| | Beta calibration | 0.072 ± 0.007 | 0.076 ± 0.007 | **0.013** ± 0.004 | 0.039 ± 0.006 | 0.051 ± 0.007 | **0.020** ± 0.003 |
| | Gaussian calibration | 0.069 ± 0.011 | 0.078 ± 0.012 | **0.016** ± 0.004 | 0.079 ± 0.015 | 0.093 ± 0.015 | **0.020** ± 0.003 |
| | Gamma calibration | 0.086 ± 0.014 | 0.101 ± 0.012 | **0.016** ± 0.003 | 0.119 ± 0.015 | 0.137 ± 0.013 | **0.020** ± 0.003 |

  - 推薦数を変化させた際，Oroginalのキャリブレーションモデルと比較して，誤差を安定的に抑えられることがわかった

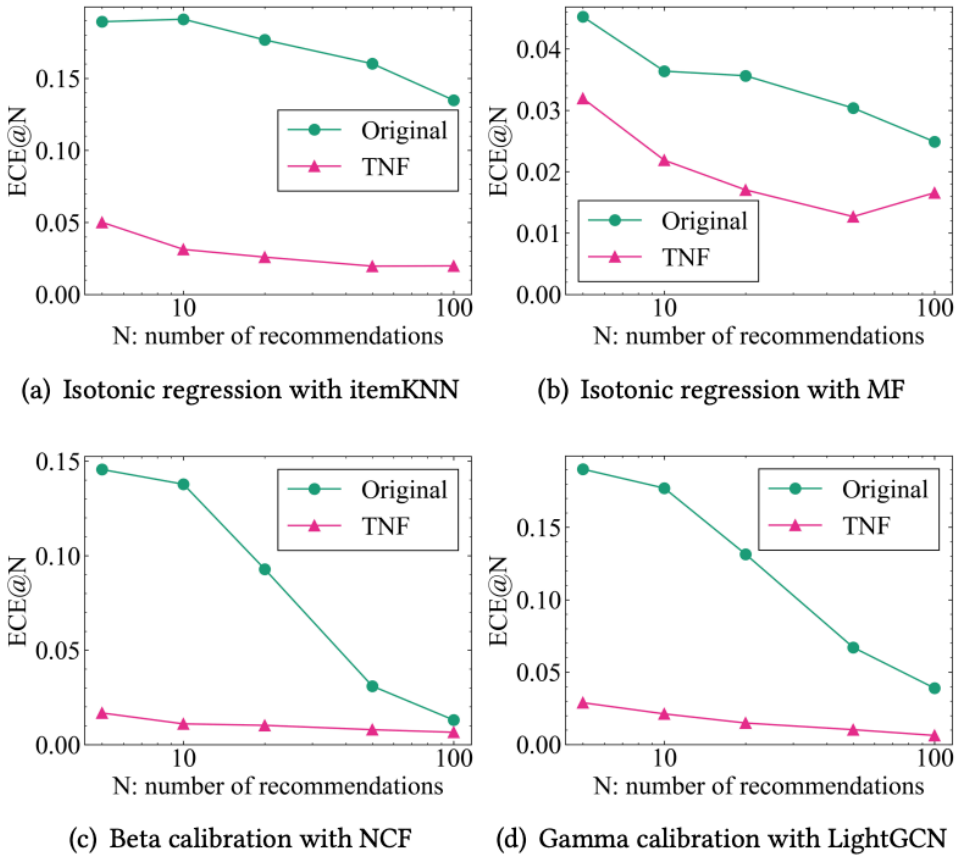



**Figure 3: ECE@N for varied number of recommendations.**

## まとめ

- Top-Nに着目してキャリブレーションの対処を行った
- 上位に重みがかかる評価手法を導入して，上位に対する評価を行った
- また上位のミスキャリブレーションを抑制するようなキャリブレーション最適化手法を導入した
- 評価予測タスク，嗜好予測タスクで既存手法を上回ることを実証した

## 所感

- 上位N件しかユーザは見ていないというユーザ目線の観点を取り入れており実用的だと感じた
- 既存手法は上位に大きな乖離があり，下位がうまくいっていたから，全体としてうまくいっているように見えていたというのは，気づかず使っているとちょっと怖いところ

- 手法が評価指標に沿ったようになっているので，他の指標で見るとどうなのかは気になる