D-003

深層学習を用いた情報推薦における初期補完値が与える推薦精度への影響
Infruence of Recommendation precision on Initial Completion Value in Information Recommendation using Deep Learning

田中 恒平 † 小林 亜樹 ‡
Kohei Tanaka Aki Kobayashi

1. はじめに

機械学習の新たなアプローチであるディープラーニング技術 (以下，DL) が話題になっている．DL は情報推薦分野においても応用され始めており，情報推薦分野で取り扱う嗜好データを DL を用いたオートエンコーダへの入力とした報告 [1][2] がある．しかし，一般的な嗜好データは大部分が欠損値であり，欠損値の存在を許容しないオートエンコーダへの入力とするためには補完する必要があるが，それについての議論はなされていない．そこで筆者らは，特定のユーザがアイテムに対し付与した評価値の平均値で補完するなどの一般的な手法で補完を行い，推薦精度について議論した [3]．

本研究では，補完値が最終的な推薦精度に及ぼす影響について検討した．

2. 推薦方法

ユーザの嗜好の学習にはニューラルネットワークの枠組みであるオートエンコーダを用いる．オートエンコーダは，入力データと出力データが等しくなるように，隠れ層の重みパラメータの更新を行い学習する．具体的な処理として，モデルの学習時には推薦対象ユーザの評価値ベクトル r が入力となる．このとき評価値ベクトル r が欠損値を含む場合には補完を行い，オートエンコーダへの入力を可能にする．隠れ層では推薦対象ユーザの特徴量が蓄積され，出力層では隠れ層の特徴量を用いて入力データが再構成される．入力データと出力層にて再構成されたデータとの誤差を算出し，誤差逆伝播法により隠れ層のパラメータを更新する．

モデルの推薦精度をテストする際には，1 人分の評価値ベクトル r_i を学習済みのモデルへ入力し，出力層にて出力された評価値ベクトル $\hat{r}_i$ を推薦評価値として取り扱う．

3. 欠損値補完法

入力となるデータに欠損値が含まれていた場合には，DL を適用することは困難である．情報推薦で用いる一般的な嗜好データは大部分が欠損しているため補完する必要がある．そこで，ユーザが評価したアイテムの平均値で補完する手法 (ユーザ平均手法)，評価値定義域内の固定値で補完する手法 (固定値手法)，評価値定義域内のランダムな値で補完を行う手法 (ランダム手法) の 3 手法で補完を行い，一般的にあまり用いられない固定値手法やランダム手法で補完した場合において，推薦精度がどの程度影響を受けるかについて検証する．

3.1. ユーザ平均手法

ユーザ平均手法は，特定のユーザがアイテムに対し付与した評価値の平均値で欠損を補完する手法である．例えば，特定のユーザ i における評価済みアイテム集合 I の評価値数が n であった場合，欠損値は $\bar{r}_i$ で補完される．

$$\bar{r}_i = \frac{1}{n}\sum_{k=1}^{n} I_{ik} \tag{1}$$

3.2. 固定値手法

固定値手法は，すべての欠損を評価値定義域内の固定値で補完する手法である．本稿では評価値定義域内の最小値，中央値，最大値で補完した．

3.3. ランダム手法

ランダム手法は，すべての欠損を評価値定義域内のランダムな値で補完する手法である．例えば評価値定義域 $R = [1,5]$ である場合，R のいずれかの値で欠損値が補完される．

4. 実験

4.1. 目的

データセット中の欠損値を補完する手法が，推薦精度に与える影響について検証する．

4.2. 条件

実験に使用するデータセットは，MovieLens と Jester の 2 種類である．MovieLens は，ユーザが映画に対して 1 から 5 の 5 段階評価をした記録を収集したデータセットであり，Jester は，ユーザがジョークに対して [-10.00,10.00] の定義域で評価をした記録を収集したデータセットである．Jester の評価値定義域は [0.00,20.00] に変更して実験を行った．

表 1: 使用データセット

	MovieLens	Jester
ユーザ数	943	73496
アイテム数	1682	100
評価値数	100000	4136360
欠損率	93.7%	43.7%

MovieLens，Jester ともに，もともとデータセットに存在する全評価値（以下，オリジナルデータ）を 5 つに等分割し 5 分割交差検証を行う．5 分割したうちの 1 つをテストデータ，残りの 4 つを学習データとして実験を行い，学習済みのモデルを 5 つ作成する．テストデータはオリジナルデータの評価値の 8 割を隠したものであり，隠された部分がオリジナルデータの評価値に近い場合には推薦精度が高いといえる．ニューラルネットワークの構築には Chainer ライブラリ [4] を

†工学院大学大学院工学研究科電気・電子工学専攻
‡工学院大学情報学部情報通信工学科

利用した．オートエンコーダは隠れ層が 1 層のみの 3 層オートエンコーダとする．隠れ層のユニット数が 10，50 の 2 パターンについて実験を行った．学習アルゴリズムは確率的勾配降下法であり，学習回数は 500 回とした．

4.3. 評価

学習済みモデルの推薦精度は RMSE で評価する．

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(r_k - \hat{r}_k)^2} \qquad (2)$$

r_k はオリジナルデータの評価値であり，$\hat{r}_k$ は推薦評価値である．RMSE は推薦精度の低さを示しており，値が小さいほうがより良い結果である．

4.4. 実験結果

ベースライン手法としてユーザ平均手法，固定値手法，ランダム手法をデータセットに適用した段階での推薦精度を示す．

表 2: ベースライン手法

	MovieLens	Jester
ユーザ平均	1.03	6.16
最小値	2.48	11.42
中央値	1.24	5.69
最大値	1.66	10.23
ランダム	1.69	7.85

図 1，図 2 は MovieLens において隠れ層ユニット数がそれぞれ 10，50 の場合の実験結果である．図 3，図 4 は Jester において隠れ層ユニット数がそれぞれ 10，50 の場合の実験結果である．

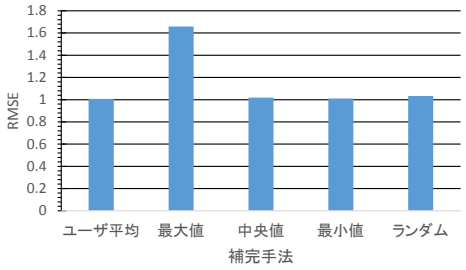

図 1: MovieLens:隠れ層ユニット数 10

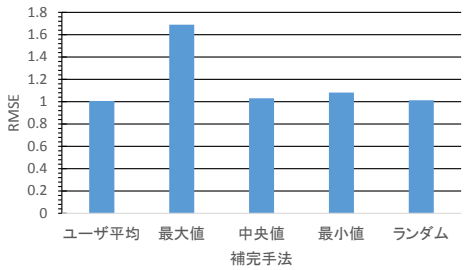

図 2: MovieLens:隠れ層ユニット数 50

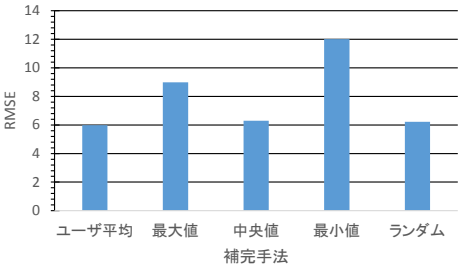

図 3: Jester:隠れ層ユニット数 10

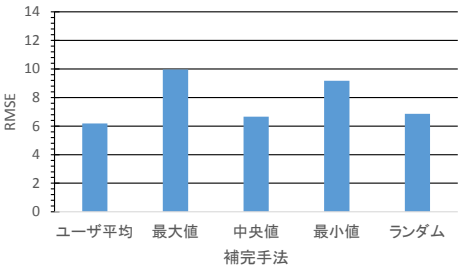

図 4: Jester:隠れ層ユニット数 50

4.5. 考察

ユーザ平均で欠損値の補完を行った場合には，他の手法と比較して推薦精度が良い傾向にある．一方，固定値手法やランダム手法では隠れ層ユニット数の増加に伴い推薦精度が下がっていることが見てとれる．MovieLens，Jester ともに隠れ層ユニット数の増加に伴って RMSE も増加しており推薦精度が低下している．これは隠れ層ユニット数が 10 の場合でも嗜好データの特徴を表現することが可能であり，50 の場合には過学習を引き起こしているものとみられる．推薦精度を向上させるためには，学習率の変更や学習アルゴリズムを変更しオートエンコーダの入出力の誤差を減らすといった方法が考えられる．

5. まとめ

本稿では欠損値を補完する手法による推薦精度の変化について検証した．その結果，評価値定義域の極端な値で補完した場合には良い推薦精度が得られなかった．一方，中庸な値で補完した場合には手法間の推薦精度の差は少なく，極端な値で補完する場合よりも推薦精度は良いので，欠損値を中庸な値で補完すればよいという知見を得た．

参考文献

[1] 川上和也，松尾豊，“Deep Collaborative Filtering: Deep Learning 技術の推薦システムへの応用” 人工知能学会全国大会論文集 28, pp.1-4, 2014.

[2] Shuiguang Deng, Longtao Huang, Guandong Xu, Xindong Wu, Zhaohui Wu, “On Deep Learning for Trust-Aware Recommendations in Social Networks” Proc.IEEE Transactions on Neural Networks and Learning Systems, pp.1-14, 2016

[3] 田中恒平，小林亜樹，“深層学習を用いた情報推薦のための欠損値補完手法” ,DEIM2016, C7-4, 2016-03-02.

[4] http://chainer.org/