4K-05

カテゴリ情報を反映させた深層学習による協調フィルタリング手法

田中恒平†　　小林亜樹†
†工学院大学工学部情報通信工学科

1　はじめに

情報推薦で用いる典型的な嗜好データは欠損値を含み，欠損したデータをディープラーニングへの入力とするためには欠損値を補完する必要がある．ディープラーニングで協調フィルタリングを行うためには，オートエンコーダの枠組みを用いた学習手法を用いる．本研究では，はじめに欠損値を補完する手法としてユーザ平均とするなどの手法を用い，推薦精度の比較を行う．さらにカテゴリ情報をデータセットに反映することで推薦精度の向上を狙う．

2　協調フィルタリング

ユーザベース協調フィルタリングの典型的な処理について述べる．ユーザベース協調フィルタリングは，アイテムを推薦するためにユーザ同士の嗜好の類似性を用いる．類似性を求めるためには，ユーザ同士が共通して評価を行ったアイテムに対し，ピアソン相関係数を用いる．さらにあるユーザがアイテムに付けた評価値 r_{ij} と，ユーザが付けた全評価値の平均値 r_i との差を求め，評価値を求めた値 r_{ij}’とすることで，ユーザによる評価値のゆらぎや偏りを緩和することができる．

$$r_{ij}' = r_{ij} - r_i \tag{1}$$

3　予備実験

3.1　目的

情報推薦で用いる典型的な嗜好データは欠損値を多く含み，ディープラーニングへの入力とするためには欠損値を補完する必要がある．そこで欠損値を定義域の中央値とする手法，ユーザが付与した評価値の平均とする手法，多重代入法の3種類で補完を行い，推薦精度の比較を行う．多重代入法は疑似完全データを複数個作成し，個別に目的とする処理，すなわちディープラーニングへの入力を行い，データの統合を行う [1]．疑似完全データを作成するために，R 言語の mi パッケージを用いる．

3.2　条件

実験は，OS が Ubuntu14.04 LTS 64bit，プロセッサが intel core i5-2400 3.10GHz，メモリが 16GiB，GPU が NVIDIA の GeForce GTX750Ti で構成された PC を用いた．ディープラーニング用ライブラリには Chainer を用いた．使用するデータセットは Movie Lens-100K[2] であり，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る．

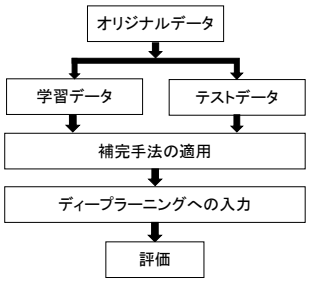

図 1: 実験概要図

表 1: Movielens-100k

項目	データ
ユーザ数	943
アイテム数	1682
評価数	100000
欠損率	93.7%

データセットはオリジナルデータの評価数の 80%を学習データとし，20%をテストデータとした．ディープラーニングは 3 層で構成し，隠れ層のユニット数は 10 から 300 の 10 刻みの 30 通りについて計測した．オートエンコーダはニューラルネットワークにおいて入力と出力が等しくなるように学習する手法であり，入力の次元数を隠れ層の次元数より少なくすることで次元圧縮が行える．入力は補完した評価値行列のユーザ一人が全アイテムに付与した評価値であり，入力次元，出力次元ともに 1682 次元である．学習アルゴリズムは確率的勾配降下法 (SGD)，モーメンタム法，Adam とし学習回数は 500 とした．推薦評価値 r_i とオリジナルデータ r_o における RMSE を算出し，推薦精度の比較を行う．

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(r_o - r_i)^2} \tag{2}$$

ここで n はオリジナルデータの評価数である．

3.3　実験結果

ベースライン手法としてユーザ平均，中央値手法を適用した段階での RMSE はそれぞれ，1.03，1.24 であった．ここでは一部の結果のみ示す．図 2，図 3 はそれぞれ欠損値を中央値手法，ユーザ平均により補完を行ったものであり，隠れ層のユニット数の増加による RMSE の変化を示している．

Collaborative Filtering using Deep Learning concerned with Item Categories.
†Kohei Tanaka　†Aki Kobayashi
†Department of Information and Communications Engineering, Faculty of Engineering, Kogakuin University

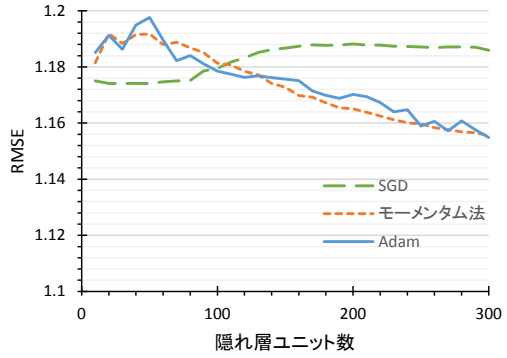

図 2: 中央値手法における RMSE

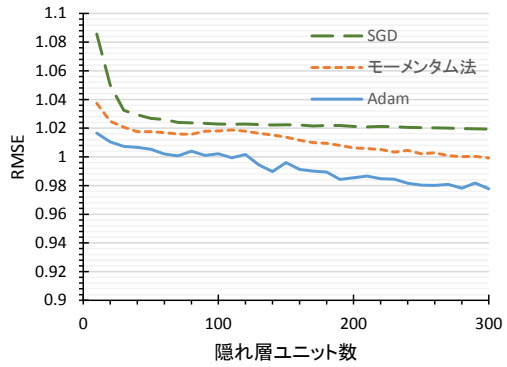

図 3: ユーザ平均における RMSE

3.4 考察

ユーザ平均の場合は，隠れ層のユニット数が多いほど RMSE は小さく，高い精度となっていることが見受けられる．これは隠れ層のユニット数を増やすことで学習データ，すなわちユーザの嗜好の特徴をより反映できることを示唆するものと考えられる．中央値手法の場合は隠れ層のユニット数が多い程，RMSE も高いことが見てとれる．中央値で補完することにより，ユーザの嗜好の特徴が失われてしまったのではないかと考える．

4 実験

4.1 手順

実験に用いる PC は予備実験と同様である．MovieLens-100K はアイテムにアクションやドラマといったカテゴリ情報が付与されており，アイテムは少なくとも 1 つ以上のカテゴリに属している．本実験では最も多く付与されているカテゴリであるドラマに属したアイテムのみをデータセットとして用いる．データセットは予備実験と同様にオリジナルデータの評価数の 80%を学習データ，20%をテストデータとした．予備実験の結果を踏まえ，欠損値はユーザ平均で補完を行い，隠れ層のユニット数は 10 から 300 の 10 刻みの 30 通りで計測した．学習アルゴリズムは確率的勾配降下法 (SGD)，モーメンタム法，Adam とし学習回数は 500 とした．

表 2: データセット

項目	データ
ユーザ数	943
アイテム数	724
評価数	39816
欠損率	94.2%

4.2 実験結果

図 4 はカテゴリ情報をデータセットに反映し，ユーザ平均で欠損値を補完した場合における隠れ層ユニット数の増加による RMSE の変化を示している．

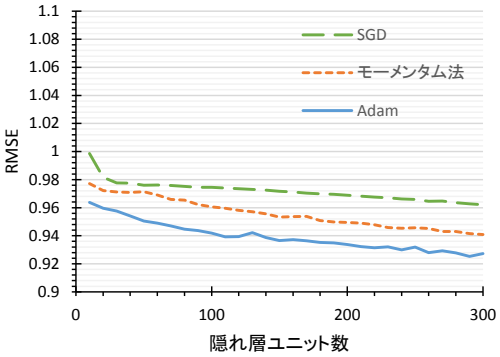

図 4: カテゴリ情報反映手法における RMSE

4.3 考察

カテゴリ情報をデータセットに反映した場合においても，隠れ層のユニット数が多い程 RMSE は小さく，高い精度となっていることが見受けられる．さらにカテゴリ情報を反映していないユーザ平均と比較すると，3 種類の学習アルゴリズムすべてにおいて精度が高くなっていることが見てとれる．

5 おわりに

本稿では欠損値補完手法毎の精度の比較を行った．さらにカテゴリ情報をデータセットに反映した場合の精度を示し，手法の有効性を示した．

参考文献

[1] Jia, Zhou., Tiejian Luo., A Novel Approach to Solve the Sparsity Problem in Collaborative Filtering., Proc.Networking, Sensing and Control (ICNSC), 2010 International Conference on, 165-170, April 2010.

[2] http://grouplens.org/datasets/movielens/