深層学習を用いた情報推薦のための欠損値補完手法

田中　恒平†　　小林　亜樹††

† 工学院大学工学部情報通信工学科　〒 163-8677 東京都新宿区西新宿 1-24-2
†† 工学院大学工学部情報通信工学科准教授　〒 163-8677 東京都新宿区西新宿 1-24-2
E-mail: †c512073@ns.kogakuin.ac.jp, ††aki@cc.kogakuin.ac.jp

あらまし　欠損値を持つ情報推薦のためのデータセットをディープラーニングへの入力とするためには，欠損値を含む行を削除するか，補完する方法がある．ユーザがアイテムに対し評価を行った嗜好データは，大部分が欠損値であるので削除する方法は適用することが難しい．そこでオリジナルデータの欠損値に対し，評価値の中央値などの方法を用いて補完を行い，疑似完全データを作成する．作成した疑似完全データを，オートエンコーダを利用したディープラーニングへの入力とし学習を行う．出力は，他のユーザの嗜好を反映した状態で疑似完全データが再構成され，オリジナルデータとの二乗平均平方根誤差 (RMSE) で評価を行う．

キーワード　深層学習, 情報推薦, 欠損値補完

1. はじめに

ディープラーニングと呼ばれる機械学習の新しい手法が話題になっている．ディープラーニングは画像認識や音声認識などの分野で研究が盛んに行われており，応用が進んでいる．応用例として，画像認識分野では全方位画像から人を検出する手法 [1] が提案されており，音声認識分野では，音声の特徴をディープラーニングにより抽出することで，感情の揺らぎによる音声認識精度の低下を防ぎ，音声認識の向上が確認されたという報告 [2] がある．ディープラーニングは多層のニューラルネットワーク (DNN) で構成されており，誤差逆伝播法により隠れ層の重みを更新することで学習を行う．誤差逆伝播法の概念は 1980 年代に登場したが，当時は隠れ層のユニット数を増やすことや，ニューラルネットワークのパラメータがランダムな初期値であることなどの要因で過学習を引き起こし，未知のデータが入力された際に性能が低下するという問題があった．過学習による性能低下の問題は，2006 年に Hinton ら [3] によるプレトレーニングという手法を用いることにより解決された．ディープラーニングについての研究が盛んに行われている理由として，先述したプレトレーニングによる過学習の防止や，計算機の性能の向上により大量のデータを用いて学習が行えるようになり，画像や音声を用いたタスクでの認識精度の向上が要因であると考えられる．

画像認識や音声認識などで応用されているディープラーニングであるが，情報推薦分野に対して応用されている例は少なく，研究があまり進んでいないのが現状である．そこで本研究は情報推薦分野の中でも，ユーザベース協調フィルタリングに対してディープラーニングを応用する．

2. 情報推薦

2.1 協調フィルタリング

協調フィルタリングは電子商取引などで応用されており，ユーザがアイテムを閲覧した時の閲覧履歴や，アイテムに付与された評価値からアイテムを推薦するアルゴリズムである．協調フィルタリングはユーザによるアイテムの閲覧履歴や評価値から，ユーザやアイテム同士の類似度を求め，アイテムを推薦する手法である．

協調フィルタリングはアイテムベース協調フィルタリングとユーザベース協調フィルタリングの 2 種類に大別される．アイテムベース協調フィルタリングは多くのユーザにより同じような評価をされているアイテムは類似していると考え，ユーザが好むアイテムと類似したアイテムを推薦する．一方ユーザベース協調フィルタリングは，嗜好が類似したユーザであるならば同じようなアイテムを好むという前提からアイテムを推薦する．

ユーザベース協調フィルタリングにおける類似度は，ユーザが共通に評価しているアイテム集合についてピアソン相関係数などを試算することで求められる．しかし，情報推薦で用いる典型的な嗜好データには，欠損値や評価値の偏りやゆらぎをもち，これらの取り扱いが問題となる．欠損値はユーザがアイテムを閲覧していない，またはアイテムを閲覧したが評価を行っていないことにより生じると考えられる．評価値のゆらぎに関しては，嗜好データに対する事前処理として評価値からユーザが付けた評価値の平均を引くことにより緩和することができる [4].

2.2 ディープラーニングによる協調フィルタリング

ディープラーニングを協調フィルタリングに応用した例として，嗜好データのゆらぎを事前処理で緩和した後に，重みを共有させたオートエンコーダを用いた多層ニューラルネットワークへの入力とすることで，従来の協調フィルタリングと比較し精度が向上するという報告があった [5]．この事前処理は，ユーザ-アイテム行列 r のユーザ i におけるアイテム j に対しての評価値 r_{ij} と，アイテム j の平均評価値 b_j との差

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

を入力とし，推薦評価値 $\tilde{r_{ij}}$ は DNN から出力された値である $\tilde{r_{ij}}'$ と，アイテム平均 b_j との和

$$\tilde{r_{ij}} = \tilde{r_{ij}}' + b_j \quad (2)$$

とする方法であった．

本研究では，嗜好データを DNN への入力とするために欠損値の補完法について議論を行い，さらにデータセットにカテゴリ情報を適用することで推薦精度の向上を目的とする．

3. 欠損値補完手法

欠損値に対して行う処理として，リストワイズ法などの欠損値を含むデータを削除する方法と，ユーザ平均などの方法で欠損値を補完する 2 通りの方法がある．しかし，嗜好データは一般に欠損率が非常に高く，9 割以上が欠損となることも珍しくないため，データを削除するリストワイズ法などの方法を適用することは難しい．そこで欠損値の補完を行う必要があり，3 種類の補完手法について言及する．

3.1 中央値手法

	i_1	i_2	i_3	i_4
u_1	4		2	5
u_2		3		1
u_3		4	2	5

→

	i_1	i_2	i_3	i_4
u_1	4	3	2	5
u_2	3	3	3	1
u_3	3	4	2	5

図 1 中央値手法

中央値手法は，評価値の定義域の中央値で欠損値を補完する手法である．例えば 1 から 5 の定義域でアイテムが評価されているならば，すべての欠損値は 3 で補完される．固定値での補完となるので，計算時間は 3 手法の中で最も短い．

3.2 ユーザ平均

ユーザ平均は，欠損値をユーザが付与した評価値の平均値で補完する手法である．図 3.2 の例では，DNN へ入力するために u_2 が評価を行っていない i_1 と i_3 に対して補完を行う必要がある．u_2 は i_1 と i_3 の 2 つのアイテムに対して評価を行っており，評価値の合計は 4 である．そのため u_2 の評価値の平均値は 2 となり，欠損値は 2 で補完を行う．

	i_1	i_2	i_3	i_4
u_1	4		2	5
u_2		3		1
u_3		4	2	5

→

	i_1	i_2	i_3	i_4
u_1	4	3.7	2	5
u_2	2	3	2	1
u_3	3.7	4	2	5

図 2 ユーザ平均

3.3 多重代入法

多重代入法は，欠損値に異なる値を代入した疑似完全データを複数個作成し，個別に目的とする処理を行いデータの統合を行う手法である．作成する疑似完全データの数が多ければ多い程補完の精度は向上するが，疑似完全データを 100 個以上作成してもかかる時間に対して得られるものが少なく，作成する疑似完全データの数は 20〜50 個程度で十分であるといわれている [6]．疑似完全データを作成する過程で用いるアルゴリズムは，マルコフ連鎖モンテカルロ法や，完全条件付指定などがある [7]．マルコフ連鎖モンテカルロ法により欠損値の補完を行い，協調フィルタリングを行った例 [8] がある．

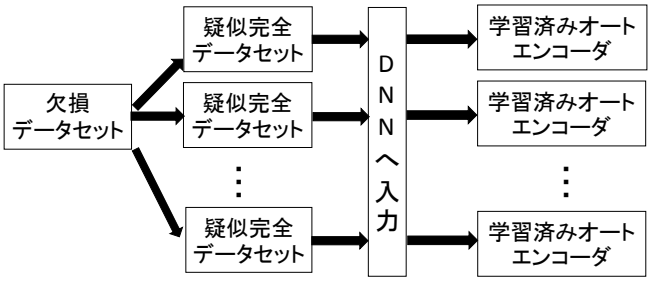

図 3 多重代入法-学習段階

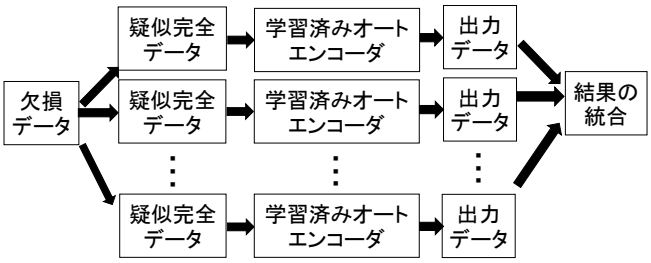

図 4 多重代入法-テスト段階

マルコフ連鎖モンテカルロ法，完全条件付指定を用いた多重代入法はそれぞれ R 言語の mi，mice パッケージで実現できる．mice パッケージを用いて嗜好データの欠損値を補完し，協調フィルタリングを行ったという例もある [9]．

本研究において多重代入法を適用する方法は，まず学習段階，テスト段階ともに mi パッケージもしくは mice パッケージにより疑似完全データを作成する．学習段階では，疑似完全学習データを個別に DNN へ入力し，学習済みオートエンコーダを作成する．テスト段階では，疑似完全テストデータを学習段階で作成した学習済みオートエンコーダに入力する．出力である推薦評価値は入力とした疑似完全データの数だけ得られ，推薦評価値の平均をとることでデータの統合を行う．

4. オートエンコーダ

図 5 情報推薦の処理

まず本研究における情報推薦の処理内容について述べる．まず推薦対象ユーザの 1 人分の嗜好データを推薦器に入力する．

この推薦器は入力された嗜好データに対して，ユーザベース協調フィルタリングの処理を行う．このとき推薦器の出力は，欠損した情報が補完された状態で出力される．図 4.1 では左側の上から 3 番目の嗜好データが欠損であり，これを入力とすると右側の出力データで*が計算された値である．この推薦器と同様の処理をオートエンコーダを用いて実現する．

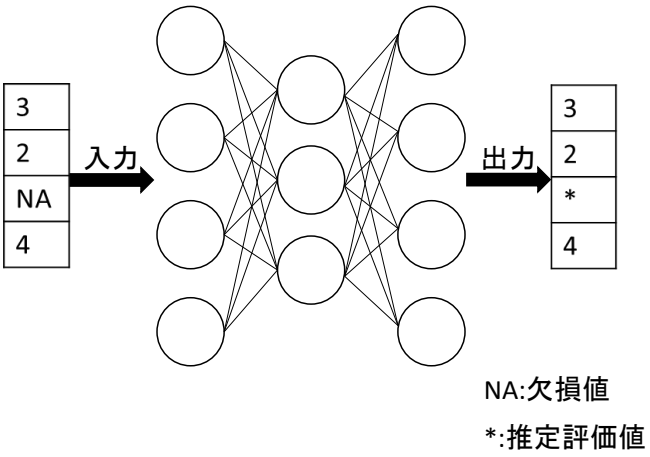

図 6　オートエンコーダ

次にオートエンコーダを用いて情報推薦を行う具体的な方法について述べる．学習段階では，様々な推薦対象ユーザの嗜好データを入力する．嗜好データは欠損値を含むデータであるため，欠損値補完手法により疑似完全データを作成し，これをオートエンコーダへの入力とする．このとき隠れ層には入力されたデータの特徴が蓄積され (エンコード)，学習を進めると隠れ層の重みに他のユーザの特徴が反映される．出力層では，隠れ層に蓄積された嗜好データの特徴を用いて入力データを再現したものが出力される (デコード)．入力データと出力データの誤差を算出し，誤差逆伝播法により隠れ層の重みパラメータを更新する．この作業を学習回数分繰り返すことにより，学習済みオートエンコーダを作成する．

テストの段階では，ユーザ 1 人分の嗜好データを学習の段階で適用した補完手法と同一の手法で補完を行い学習済みオートエンコーダに入力すると，出力層では隠れ層に蓄積された様々なユーザの特徴を用いて入力データを再現したものが出力される．これが協調フィルタリングと同様の処理となり，出力層における出力を推薦評価値として取り扱う．

オートエンコーダにおける隠れ層のユニット数を，入力，出力次元以上の値にしてしまうとネットワークが恒等写像を学習してしまい無意味なものとなる．そのため隠れ層のユニット数は，入力データの特徴を隠れ層に蓄積するために，入力，出力次元数よりも少なくする必要がある．

エンコードは x を入力層における入力とすると (4.1) 式のようになり，デコードは (4.2) 式のようになる．

$$y = f(Wx + b) \tag{3}$$

$$z = f(W'y + b') \tag{4}$$

ここで W，W' はそれぞれエンコード，デコードにおける重みであり，本研究では $W' = W^T$ のように制約を付けた．

オートエンコーダの拡張として入力データにノイズを付加し，ノイズを付加させる前のデータを出力とするように学習させるデノイジングオートエンコーダなどがあるが，本研究ではオートエンコーダについてのみ実験を行った．

欠損値の補完を行う際に学習段階とテストの段階で異なる欠損値補完法を適用することも可能であるが，本研究では学習段階とテストの段階で適用する補完法は同一とする．

5. 実　　験

5.1 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，DNN 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 (中央値手法)，ユーザ平均，多重代入法の 3 種類で補完を行う．学習段階とテスト段階で補完手法を同一とした場合における推薦精度の比較を行うことが目的となる．

5.2 条　　件

実験は表 1 のように構成された PC で行った．ディープラーニング用ライブラリには Chainer [10] を用いた．使用するデータセットは MovieLens-100K [11] 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．

表 1　実 験 環 境

項目	データ
OS	Ubuntu14.04 LTS 64bit
プロセッサ	intel core i5-2400 3.10GHz
メモリ	16GiB
GPU	NVIDIA GeForce GTX750Ti

表 2　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%

MovieLens-100K データ (以下，オリジナルデータ) における評価数の 80%を学習データ，20%をテストデータとした．DNN は 3 層で構成し，隠れ層のユニット数は 10 から 300 の範囲で 10 ずつ増加させ，30 通りについて実験を行った．オートエンコーダの重みは $W' = W^T$ のように制約をかけ，活性化関数はシグモイド関数とした．欠損値の補完は，嗜好データを入力しオートエンコーダを学習させる段階と，学習済みオートエンコーダに嗜好データを入力しテストを行う段階それぞれにおいて行う必要がある．本研究ではオートエンコーダを学習させる段階とテストの段階での補完法は，異なる補完法を用いるのではなく一致させて実験を行った．補完後のユーザ 1 人分の嗜好データをオートエンコーダへの入力とし，入力次元，出力次元ともに 1682 次元である．

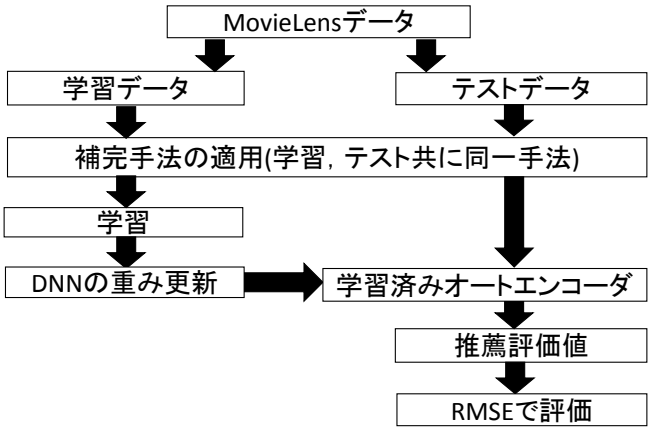

図 7　実験概要図

5.3　評　　価

推薦精度の評価は，まずテストデータに対して学習データに適用した補完法と同一の手法で補完を行い，学習済みオートエンコーダへ入力し推薦評価値を得る．テストデータはオリジナルデータの評価値の 80%を隠したものであり，隠された部分の推薦評価値がオリジナルデータの評価値に近ければ推薦精度が高いといえる．評価指標には 2 乗平均平方根誤差 (RMSE) を用いる．

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(r_{ok} - r_{ik})^2} \qquad (5)$$

r_i は推薦評価値，r_o はオリジナルデータにおける評価値である．n はオリジナルデータの評価値数を示している．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．

5.4　学習アルゴリズム

学習アルゴリズムは誤差関数を最小化し隠れ層のパラメータを更新することが目的となる．学習アルゴリズムは，学習のステップ幅である学習率の値が小さすぎると延々として学習が進まず，学習率の値が大きすぎると最適な値を飛び越えてしまい収束しない恐れがある．そのため勾配が大きい場合には学習率を大きく，勾配が小さい，すなわち最小解や局所解付近では学習率を小さくすると効率良く学習が行える．

5.4.1　確率的勾配降下法 (SGD)

SGD はまずデータセット中から適当な数だけデータを取り出し，誤差関数の勾配を求める．そして，勾配の向きに勾配と学習率の積の分だけパラメータを更新する．

5.4.2　モーメンタム法

SGD の最小解あるいは局所解にたどり着くまでの時間 (以下，収束時間) を短縮するために，モーメンタム法が用いられることもある．モーメンタム法は勾配の符号の変化が少ない場合には学習の速度が速く，符号の変化が多い場合には学習の速度を落とす手法である．

5.4.3　Adam

Adam は学習率を自動で調整する手法であり，収束時間が SGD やモーメンタム法と比較して速い．さらに SGD やモーメンタム法と比較してより良い局所解を得られる傾向がある．

5.5　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 を表 3 に示す．

表 3　ベースライン手法

補完手法	データ
中央値手法	1.24
ユーザ平均	1.03

多重代入法による欠損値の補完には R 言語の mi パッケージ，mice パッケージを用いた．mi パッケージによる多重代入法は，所要時間が評価値行列のサイズに大きく依存する．そのため MovieLens-100K データに対し mi パッケージをそのまま適用すると計算量が多大になり，本実験環境では 3 週間かかっても処理が完了しなかったため，打ち切った．

一方，mice パッケージを MovieLens-100K データに適用すると，疑似完全データを 4 個作成した場合の所要時間は 10 日程となった．しかし，作成した疑似完全データは 72.1%が欠損値のままであり，完全に欠損値が補完されていなかった．

以上の理由から，ここでは中央値手法とユーザ平均により欠損値の補完を行った結果のみ示す．図 5.2，図 5.3 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．

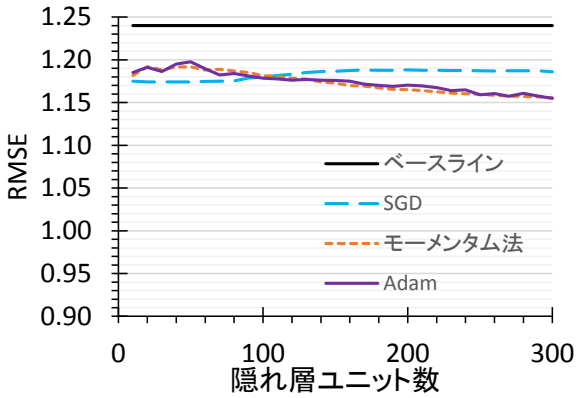

図 8　中央値手法における RMSE

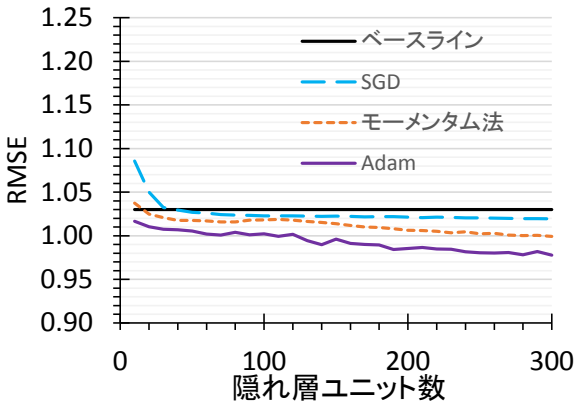

図 9　ユーザ平均における RMSE

5.6 考　　察

図 5.3 のユーザ平均で補完を行った場合の結果は，全ての学習アルゴリズムにおいて，隠れ層のユニット数が多いほど RMSE は小さくなり減少傾向を示している．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．

図 5.2 の中央値手法で補完を行った場合の結果は，学習アルゴリズムがモーメンタム法，Adam の場合には隠れ層のユニット数が多いほど RMSE は小さくなり減少傾向を示している．一方，学習アルゴリズムが SGD の場合には，隠れ層のユニット数が多いほど RMSE も大きくなり推薦精度が低くなっている．さらに図 5.3 のユーザ平均で補完を行った場合の結果と比較して，中央値手法は RMSE の値が全ての学習アルゴリズムにおいて高く，推薦精度が低いという結果になった．これは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．

6. カテゴリ情報の反映

6.1 目　　的

データセットにカテゴリ情報を付与することにより，推薦精度を向上させる．さらにデータセットの量や，アイテム数を減らすことによる推薦精度の変化についても考察を行う．

6.2 手　　順

実験に用いる PC は 5 章の実験で用いたものと同様である．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．本実験ではドラマ，コメディ，アクションの 3 種類に属したアイテムのみをデータセットとして用い，カテゴリごとに実験を行った．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，20%をテストデータとした．5 章の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 の範囲で 10 ずつ増加させ，30 通りについて実験を行った．アクションカテゴリのみアイテム数が 251 であるため，隠れ層のユニット数は 10 から 250 までの範囲で実験を行った．学習アルゴリズムは SGD，モーメンタム法，Adam とし学習回数は 500 とした．

表 4　実験データセット

項目	ドラマ	コメディ	アクション
ユーザ数	943		
アイテム数	724	505	251
評価数	39816	29832	25589
欠損率	94.2%	93.7%	89.2%

6.3 実 験 結 果

図 6.1，6.2，6.3 はそれぞれドラマ，コメディ，アクションカテゴリのみをデータセットとして用い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．

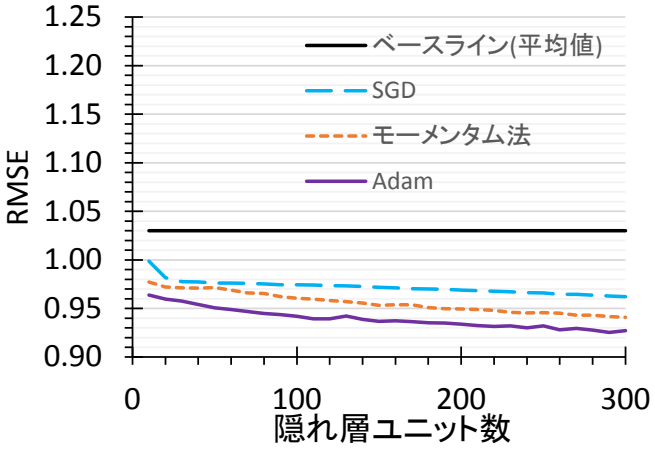

図 10　ドラマカテゴリにおける RMSE

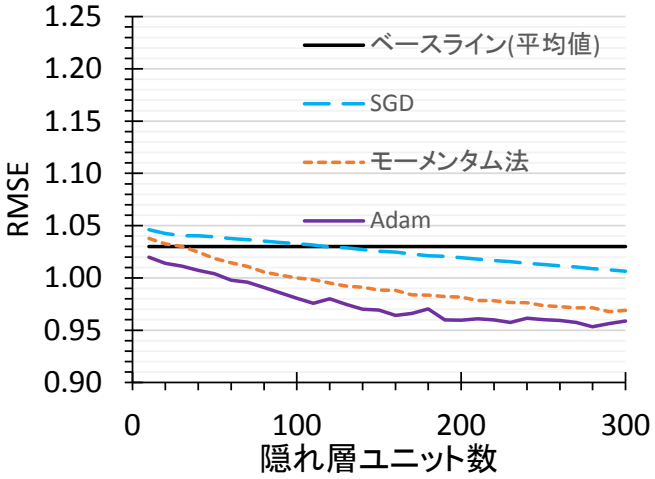

図 11　コメディカテゴリにおける RMSE

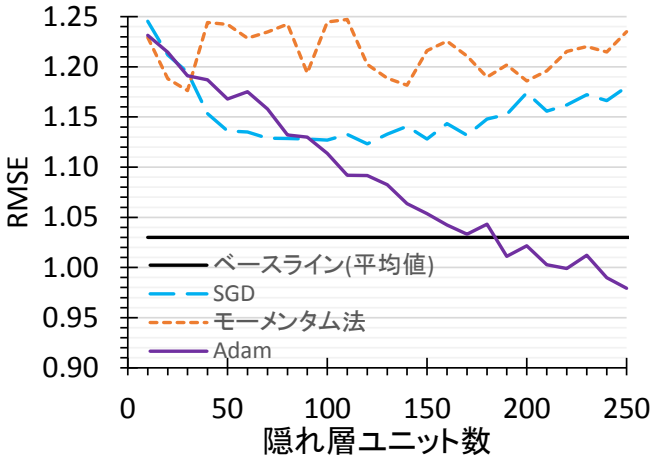

図 12　アクションカテゴリにおける RMSE

6.4 考　　察

ドラマカテゴリ，コメディカテゴリを反映した場合においては，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．一方アクションカテゴリを反映した場合には Adam のみ隠れ層ユニット数の増加による RMSE の減少傾向を示し，SGD 法，モーメンタム法に関しては RMSE の減少傾向を示していないことが見てとれる．

オートエンコーダは隠れ層のユニット数を入力次元，出力次元と等しくした場合恒等写像を学習してしまう．アクションカテゴリの場合はアイテムの数がドラマカテゴリやコメディカテゴリと比較して少なく，隠れ層のユニット数を増やしていくにつれ恒等写像に近い学習をしたために RMSE が減少傾向を示さなかったのではないかと考える．

評価値数はドラマカテゴリが最も多く，次いでコメディカテゴリ，アクションカテゴリが最も少ない．実験結果はドラマカテゴリ，コメディカテゴリ，アクションカテゴリの順に良いということから，データセットの量が推薦精度に影響を及ぼしていることも考えられる．さらにカテゴリ情報を反映していないユーザ平均の実験結果と比較すると，ドラマカテゴリにおいては 3 種類の学習アルゴリズムすべてにおいて RMSE の値が小さく，推薦精度が高くなっていることが見てとれる．

この結果から，データセットに制限を付けて協調フィルタリングを行うほうが，ユーザの嗜好を捉えた推薦ができているのではないかと考える．しかし，ディープラーニング技術を応用した場合には，アクションカテゴリのような小規模なデータセットを用いた場合は学習がうまく行えず，推薦精度が良くはならないという知見を得た．

7. ま と め

本論文では，はじめに欠損値の補完手法毎の精度の比較を行い，ユーザが付与した評価値の平均で補完をする手法が良い推薦精度を示した．さらにカテゴリ情報をデータセットに反映した場合の推薦精度を示し，ある特定のカテゴリに属したアイテムのみをデータセットとすることで，推薦精度が向上することを示した．今後は多重代入法を嗜好データに適用する方法を考え，他の補完手法と比較を行う予定である．さらに matrix factorization などの既存手法との推薦精度の比較も行う予定である．

文　　献

[1] 浅沼仁, 川本一彦, 岡本一志, “Deep Convolutional Neural Network による全方位画像からの人検出” 研究報告コンピュータビジョンとイメージメディア（CVIM） 2015-CVIM-195(59), 1-4, 2015-01-15.

[2] 向原康平, サクリアニ サクティ, グラム ニュービック, 戸田智基, 中村哲, “ボトルネック特徴量を用いた感情音声の認” 研究報告音声言語情報処理（SLP）, 2015-SLP-107(2), 1-6, 2015-07-09.

[3] Geoffrey E. Hinton, Simon Osindero “A fast learning algorithm for deep belief nets.” Neural Computation, 18, pp 1527-1554, 2006.

[4] 神嶌敏弘, “推薦システムのアルゴリズム (2)” 人工知能学会誌 23(1), 89-103, 2008-01-01.

[5] 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, 1-4, 2014.

[6] 高橋将宜, 伊藤孝之, “大規模経済系データにおける様々な多重代入法アルゴリズムの検証” 2013 年度科学研究費シンポジウム～統計科学の新展開～（金沢大学） 2013 年 11 月 29 日.

[7] 高橋将宜, 伊藤孝之, “様々な多重代入法アルゴリズムの比較：大規模経済系データを用いた分析” 統計研究彙報 (71), 39-81, 2014-03.

[8] 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.

[9] Xiaoyuan Su, Taghi M. Khoshgoftaar, Russell Greiner, “A Mixture Imputation-Boosted Collaborative Filter” Proceedings of the Twenty-First International FLAIRS Conference, 312-317, 2008.

[10] http://chainer.org/

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