機械学習を用いた映画・アニメレビューの数値予測モデル

竹内 佑汰(17X4039)　指導教員 五島 洋行

1. はじめに

本研究では,消費者のニーズに合った商品を提案するシステムの構築を目的とし,機械学習を用いた映画･アニメレビューの数値予測モデルを提案する．予測精度向上のため，提案するモデルのパラメータの最適化を行うとともに，他のモデルと比較し,提案手法が予測精度において優位であるか検証する．

2. 実験で用いる数値予測モデル

本研究では，4 つの異なるモデルを用いて数値予測を行い,提案するモデルの有用性を確認する．

2.1 提案モデル

機械学習は，コンピューターにデータを反復学習させ，それに基づいて未知のものを予測，判断する技術である．本研究では，機械学習を用いた数値予測モデルの提案を行う．

本モデルでは，Matrix Factorization[1]の手法を用いて予測評価点行列をユーザーの特徴行列$P \in \mathbb{R}^{k\times m}$とアイテムの特徴行列$Q \in \mathbb{R}^{k\times n}$に行列分解し，実測値と予測値の誤差が最小となる目的関数を解く．このときmをユーザ数，nをアイテム数とする．ユーザーの特徴行列を$\boldsymbol{p}_i \in \mathbb{R}^k$，アイテムの特徴行列を$\boldsymbol{q}_u \in \mathbb{R}^k$，正規化項のパラメータを$\lambda_P$，$\lambda_Q$とするとき，目的関数は次のように表される．

$$\text{Minimize} \sum_{(u,i)\in R} (r_{u,i} - \boldsymbol{p}_u^T\boldsymbol{q}_i)^2 + \lambda_P\|\boldsymbol{p}_u\|_F^2 + \lambda_Q\|\boldsymbol{p}_i\|_F^2 \quad (1)$$

最適化問題を解くためのアルゴリズムとして，本研究では確率的勾配降下法(stochastic gradient descent，以下 SGD)[1]を採用する．SGD では，式(1)の勾配を全てのデータに対して調べるのではなく，(u,i)の候補をランダムに 1 つ取り出して誤差を計算し,パラメータの更新を行う．取り出した$r_{u,i}$によって，目的関数は以下の式によって更新される．

$$\left(r_{u,i} - \boldsymbol{p}_u^T\boldsymbol{q}_i\right)^2 + \lambda_P\boldsymbol{p}_u^T\boldsymbol{q}_i + \lambda_Q\boldsymbol{p}_i^T\boldsymbol{q}_u \quad (2)$$

$\boldsymbol{q}_i$の勾配を計算した後，αを機械学習の学習率とするとき各変数は以下のようにして更新される．

$$\boldsymbol{p}_u \leftarrow \boldsymbol{p}_u + \alpha\left(e_{u,i}\boldsymbol{q}_i - \lambda_P\boldsymbol{p}_u\right),$$

$$\boldsymbol{q}_i \leftarrow \boldsymbol{q}_i + \alpha\left(e_{u,i}\boldsymbol{p}_u - \lambda_Q\boldsymbol{q}_i\right),$$

where

$$e_{u,i} = r_{u,i} - \boldsymbol{p}_u^T\boldsymbol{q}_i \quad (3)$$

以上の手法を機械学習によって実現する.訓練用モデルを学習させ,テスト用データで数値予測を行いながら予測精度を高めていく．

2.2 協調フィルタリングモデル

このモデルは情報推薦システムの手法として広く用いられている[2]．協調フィルタリングモデルによる数値予測は以下の二段階のステップで行われる．

Step1. 類似度の計算：ユーザー間の嗜好パターンを定量的に表す.

Step2. 嗜好の予測：ユーザーが知らないアイテムに対して,他のユーザーとの類似度をもとにどの程度好むかを予測する.

2.3 ランダムモデル

ランダムモデルは,訓練用データを評価値の範囲内で全て乱数値として与えたモデルである．予測値をすべて乱数で求めるため，予測精度は低くなると考えられる．

2.4 平均値モデル

平均値モデルは,予測値をすべてのデータの平均値として与えたモデルである.正確な数値予測は望めない一方で,比較的高い予測精度を得ることができると考えられる．

3. 使用データと実験環境

本研究の実験においてアルゴリズムを実装し，精度評価を行うためのデータとして，Group Lens 社が提供する映画レビューデータと，MyAnimeList 社が提供するアニメレビューデータを用いる.それぞれに格納されたデータのうち,アイテムを鑑賞した人物のタグである userId，視聴したアイテムのタグである itemId，評価したアイテムのレビュー値である ratings の 3 つの列のみを取り出して用いる．

また,本研究のコンピュータの実験環境は表1のとおりである.

4. 実験内容

本研究では，提案するアルゴリズムの精度を高めるためにパラメータの最適化を行う実験と,最適化されたパラメータをもとにアルゴリズムの実装を行い,他のアルゴリズムとともに精度評価を行い比較する実験の 2 工程を実施する.

4.1 提案手法の精度向上

機械学習では,パラメータの設定がアルゴリズムの精度を左右する．しかしながら，パラメータの決め方は確立されていないこと,タスクごとにパラメータが異なることから,実験によってパラメータを調整する必要がある.本研究の提案モデルで調整するパラメータは以下の３つである.

① 行列P, Qの次元k
② 正規化項のパラメータλ_P, λ_Q
③ エポック数

4.2 モデルの精度評価

モデルの精度評価をすることで,モデルの有用性を確認することができる．本節では，モデルの精度評価に RMSE と MSE を採用する．r_jをユーザjの実測値，$\hat{r}_j$をユーザjの予測値，Nをデータ数とすると､評価関数は以下のように表される.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{r}_j - r_j\right)^2}{N}} \qquad (4)$$

$$MSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{r}_j - r_j\right)^2}{N}} \qquad (5)$$

5. 実験結果

パラメータの最適化を行った結果は表 2 のとおりである.

表 1 コンピューターの実験環境

OS	macOS Big Sur Ver.11.6
メモリ	16.0GB
Core	8
プログラム言語	R Ver. 4.1.1

表 2 各パラメータの最適値

	k	λ_P	λ_Q	エポック数
値	200	0.1	0.1	50

表 3 映画レビューデータの数値予測結果

アルゴリズム	RMSE	MAE
ランダム	1.932	1.575
平均値	1.054	0.834
協調フィルタリング	1.072	0.829
提案モデル	0.849	0.656

表 4 アニメレビューデータの数値予測結果

アルゴリズム	RMSE	MAE
ランダム	4.218	3.427
平均値	2.264	1.813
協調フィルタリング	2.036	1.535
提案モデル	1.830	1.258

また，映画・アニメレビューデータの数値予測の実験結果はそれぞれ表 3，表 4 のとおりである．実験の結果，映画・アニメレビューデータのどちらにおいても,提案手法の予測精度が最も優れていることが分かる．また，RMSE，MAE のどちらの評価関数においても本研究の提案するモデルの有用性を確認することができた.

6. おわりに

本研究では,機械学習を用いた数値予測モデルを提案し,パラメータの最適化を行った提案手法と，その他 3 つのモデルを用いて，レビュー値の数値予測実験を行った.数値予測を行う対象として，映画とアニメレビューデータの 2 種類を扱い,どちらのデータの適応においても提案手法の有用性が確認できた.

本研究は,今後レビュー値をもとにユーザーに対して最適なサービスを提供する際の 1 つの解決策となると考えられる.

参考文献

[1] Zhuang，W-S Chin，Y-C Juan，and C-J Lin，“A fast parallel SGD for matrix factorization in shared memory systems”，Publication:RecSys ‘13: Proceedings of the 7th ACM Conference on Recommender Systems，pp.249-256，2013.

[2] 神嶌敏弘，“推薦システムのアルゴリズム(3)”人工知能学会誌，vol.23，no.2，pp.248-263，2008.

令和 3 年度卒業研究論文

機械学習を用いた映画・アニメレビューの数値予測モデル

法政大学　理工学部　経営システム工学科

経営数理工学研究室

17X4039 竹内　佑汰

指導教員　五島　洋行　教授

学科名	経営システム工	学籍番号	17X4039
申請者氏名		竹内佑汰	
指導教員氏名		五島洋行	

論文要旨

論文題目	機械学習を用いた映画・アニメレビューレビューの数値予測モデル

企業が消費者に合わせたマーケティングを行う際に活用する指標として，レビューデータが挙げられる．レビューデータの分析や予測によって，企業の機会損失を減らし，消費者の傾向やニーズに即したサービスの提案が可能となる．

本研究では，機械学習を用いた映画・アニメレビューの数値予測モデルの提案を行う．数値予測モデルを実装する上で，予測精度を可能な限り高め，その結果を生かした提案は，顧客のサービス満足度を高めることに繋がる．

そこで，本研究では，複数の数値予測モデルを実装し，その有効性を確認する．実験の結果，機械学習の手法を用いた行列分解を行うモデルにおいて最も高い予測精度が算出されることが分かった．

目次

第 1 章 序章

1.1 研究背景と目的

近年，通信技術の発展によって，インターネット上に大量の情報が発信されるようになった．図 1 から分かるように，世界のインターネット普及率は年々増加傾向にあり，今後更にインターネットの利用者は増えると予想される．インターネットの普及に伴い，人々は様々な情報の中から自分自身に合った情報を得ることができるようになった．一方で，自分が望む情報が存在するとしても，大量の情報の中に埋もれてしまう情報過多の問題が発生してしまっているというのが現代の抱える問題である． そのため，情報を発信する立場においては，インターネットの利用者の特徴に応じて提供する内容を変えることで，利用者にとって有用な情報を的確に伝えることが求められるようになっている．

人々がインターネットでサービスの良し悪しを評価する際に活用する最も代表的な媒体として挙げられるのが，レビューサイトである．1990 年代後半にレビューを投稿できるウェブサイトが登場して以降，様々な製品やサービスに対し，専門家ではない一般の消費者がインターネット上にレビューを書き込むようになった．それによってサービスの供給者はレビューサイトの動向を注視しなければならない状況となっている．そして，レビューサイトは消費者の購買力に影響をもつことが確認されており[1]，インターネットの利便性向上とともに今後，その重要性が高まっていくと予想される．レビューサイトは主に，(1)レビュー値，(2)レビューコメントによって構成されるが，本論文では(1)を扱う．

これらの点を踏まえ，本研究では，ユーザーのニーズに合った商品を提案する情報推薦システムの構築の事例として，機械学習を用いた映画・アニメレビューの数値予測モデルを提案する．予測精度向上のため，提案するモデルのパラメータの最適化を行うとともに，他のモデルと比較し，提案手法が予測精度において優位であるか検証する．これによって，本研究がインターネットサービスの供給者がユーザーのニーズに合った商品を高い精度で推薦するための 1 つの解決策となることを目的とする．

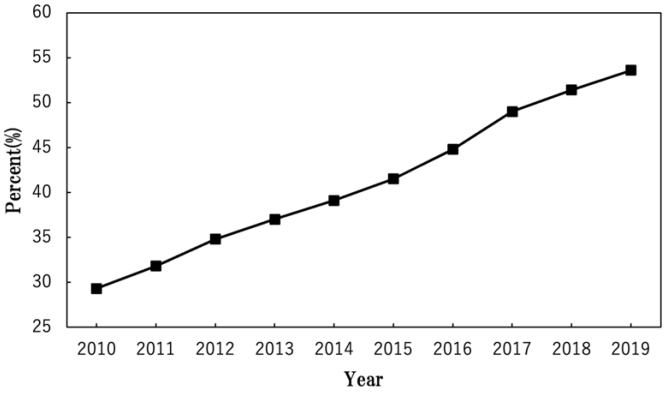

図 1 世界のインターネット普及率の推移

1.2 本論文の構成

本論文は全 7 章で構成されている．

第 2 章では，機械学習について述べる．

第 3 章では，関連・先行事例について述べる．

第 4 章では，使用データと前提条件について述べる．

第 5 章では，本論文の実験内容について述べる．

第 6 章では，5 章をもとに行った実験の結果と考察を述べる．

第 7 章では，本論文の結論を述べる．

第 2 章 機械学習

本章では，本研究のモデルで扱う機械学習の概要と，機械学習の種類について述べる．

2.1 機械学習とは

機械学習とは，コンピューターにデータを反復学習させ，それに基づいて未知のものを予測，判断する技術である．近年のコンピューティング・テクノロジーの発展により，機械学習のアルゴリズムは膨大なデータに対しても超高速で，自動的に，そして何度も反復して適用できるようになった．機械学習の活用は，これまで人々の勘や経験に基づいて判断されてきた事象などを定量的に判断することを可能にする．

2.2 機械学習の分類について

機械学習は大きく 3 つの分類に分けることができる．

① 教師あり学習
② 教師なし学習
③ 強化学習

本論文では①の手法を用いて機械学習を行う．①と②の違いを明確に示すため，③の説明は省略し，以下それぞれの持つ特徴について述べる．

① 教師あり学習

教師あり学習とは，訓練データによる「学習」とテストデータによる「予測」の 2 段階に分けて機械学習を行う手法である．学習では，正解のラベルがついた訓練データから，そのデータのパターンやルールを学習させ，モデルとして出力する．その後，正解が分からないテストデータを用意し，学習で得たモデルを適応させ，予測を行う．教師あり学習は主に，過去のデータをもとに将来起こりそうな事象を予め推測することに用いられる．代表的な手法として，分類や回帰などが挙げられる．本研究ではこの手法を用いて機械学習を行う．

② 教師なし学習

教師なし学習とは，正解のラベルがついた訓練データを与えることなく入力データを学習させる手法である．大量のデータを読み込むことによってコンピューター自身がデータの持つ特徴を導き出し，結果を出力する．教師なし学習は主に，情報の集約やグループ分けに用いられる．代表的な手法として近傍法マッピング，特異値分解などがある．

2.3 機械学習の過学習について

本節では，機械学習を行う際に留意すべきである過学習について述べる．過学習とは，学習データを忠実に再現してしまい，未知のデータに対する汎化性がなくなってしまった状態を指す[2]．過学習が生じてしまっているかどうかを判断する際には訓練用データに対する誤差（訓練誤差）と，テスト用データに対する誤差（汎化誤差）を見ることで予測を立てることが可能となる．

第 3 章 関連知識・先行事例

本章では，本論文の内容における関連知識と，先行事例について述べる．

3.1 情報推薦システム

本節では，本研究の関連知識として扱う情報推薦システムについて述べる．

情報推薦システムとは，利用者に有用と思われる対象，情報，または商品などを選び出し，それを利用者の目的に合わせた形で提示するシステム[2]である．インターネット上の大量の情報の中から有益な情報を抽出し，利用者にとって最適な提案を行う上で，情報推薦システムの構築は一つの解決策になりうるだろう．

すなわち，本研究は映画とアニメのレビュー値をもとに視聴者の趣向に合ったコンテンツを提案する情報推薦システムの構築だといえるであろう．

3.2 神嶌の研究

本節では，本研究の先行事例として神嶌の研究[3]について述べる．神嶌の研究では，利用者の嗜好に合う商品の提案を行うための手法として，先に述べた情報推薦システムのアルゴリズムの構築について記述している．そのアルゴリズムのうちの一つとして提案されているのが，利用者間型メモリベース協調フィルタリングである．次節では，利用者間型メモリベース協調フィルタリングを用いて数値予測を行う例を紹介する．

① 利用者間型メモリベース協調フィルタリング

利用者間型メモリベース協調フィルタリングの有名な手法である GroupLens の方法について述べる．この手法は，情報推薦システムのアルゴリズムにおける古典的な手法として広く用いられており，関連研究も多く行われている．

GroupLens の方法による推薦は，次の二段階で行われる．

Step1. 類似度の計算：ユーザー間の嗜好パターンを定量的に表したものである．
Step2. 嗜好の予測：ユーザーが知らないアイテムに対して，他のユーザーとの類似度をもとにどの程度好むかを予測する．

上記の 2 つのステップに対して説明をする前に，記号の定義を行う．

$\mathcal{X} = \{1, \ldots, n\}$ ：n人のユーザー全体の集合
$\mathcal{Y} = \{1, \ldots, m\}$：$m$種類の全アイテムの集合
$\mathcal{R}$ ：評価値の集合
r_{xy} ： 利用者$x \in \mathcal{X}$の，アイテム$y \in \mathcal{Y}$への評価値
(評価済みであれば定義域 $\mathcal{R}$ のいずれかの値をとり，未評価であれば欠損値
$\perp$ をとる)
$\mathbb{R}$ ：利用者$x \in \mathcal{X}$の，アイテム$y \in \mathcal{Y}$への評価値 r_{xy}を要素とする行列

まずは Step1.について考える．ユーザーを添え字aで表すとすると，r_{ay}はユーザーのアイテムyに対する評価値となる．また，ユーザーxが評価済みのアイテムの集合を $\mathcal{Y}_x = \{y \mid y \in \mathcal{Y},\ r_{xy} \neq \perp\}$ と表す．ユーザーaとユーザーxの類似度は Pearson 相関によって以下のように表される．

$$\rho_{ax} = \frac{\sum_{y \in \mathcal{Y}_{ax}} (r_{ay} - \bar{r}'_a)(r_{xy} - \bar{r}'_x)}{\sqrt{\sum_{y \in \mathcal{Y}_{ax}} (r_{ay} - \bar{r}'_a)^2} \sqrt{\sum_{y \in \mathcal{Y}_{ax}} (r_{xy} - \bar{r}'_x)^2}} \tag{1}$$

ただし，$\mathcal{Y}_{ax}$はユーザーaとxが共通に評価したアイテムの集合で，$\bar{r}'_x = \sum_{y \in \mathcal{Y}_{ax}} r_{xy} / |\mathcal{Y}_{ax}|$である．なお，$|\mathcal{Y}_{ax}| \leq 1$すなわちユーザー$a$とユーザー$x$がともに評価を行ったアイテムが 1 つ以下である場合は，Pearson 相関は計算できないので$\rho_{ax} = 0$となる．

次に，Step 2 について考える．アイテム$y \notin \mathcal{Y}_a$の評価式は，式(1)の類似度で重み付けした各ユーザーのアイテムyに対する評価値の加重平均で予測する．

$$\hat{r}_{ay} = \bar{r}_a + \frac{\sum_{x \in \mathcal{X}_y} \rho_{ax} (r_{xy} - \bar{r}_x)}{\sum_{x \in \mathcal{X}_y} |\rho_{ax}|} \tag{2}$$

ただし，$\mathcal{X}_y$はアイテムyを評価済みの利用者の集合であり，$\bar{r}_x$はユーザーxの全ての評価アイテムに対する平均評価値$\bar{r}_x = \sum_{y \in \mathcal{Y}_x} r_{xy} / |\mathcal{Y}_x|$である．

② 利用者間型メモリベース協調フィルタリングの具体例

ここでは，飲食店の推薦の例を考える．飲食店を決める際，自分の食事の好みと似た複数人の意見を取り入れることによって，食事に満足できる可能性を高めることができるだろう．

表 1 はある大学生の食べ物に対する評価行列$\mathbb{R}$の例である．4 人のユーザーは各行に対応しており，4 種類のアイテム（食べ物）に対応する．評価値は 3 段階の採点法とするため，$\mathcal{R} = \{1,2,3\}$と定義できる．ユーザーを源田，すなわちa=2 として，2：源田の 1：とんかつへの推定評価値$\hat{r}_{2,1}$を求めることにする．まず，式(1)の相関係数を求める．とんかつを評価済みのユーザー，すなわち$\mathcal{X}_1$に含まれる各ユーザー間の相関係数を計算する．

表 1 評価行列 R の例

	1:とんかつ	2:寿司	3:ステーキ	4:餃子
1:山川	1	3	⊥	3
2:源田	⊥	1	3	⊥
3:高橋	2	1	3	1
4:今井	1	3	2	⊥

1：山川，3：高橋，4：今井の 3 人ともとんかつを評価済みであることから，$\mathcal{X}_1$ ={1,3,4}の各ユーザーとの相関係数を求める．2：源田と 1：山川の相関係数は，ともに評価しているアイテムが 2：寿司だけで 1 個以下なので$\rho_{2,1}$=0 となる．次に，2：源田と 3：高橋の間の相関係数を計算する．この 2 人のユーザーがともに評価しているアイテムは 2：寿司と 3：ステーキなので$\mathcal{Y}_{2,3}$の平均評価値はそれぞれ

$$\bar{r}_2' = (\sum_{y=2,3} r_{2,y})/2 = (1+3)/2 = 2$$

$$\bar{r}_3' = (\sum_{y=2,3} r_{3,y})/2 = (1+3)/2 = 2$$

となり，相関係数は次式になる．

$$\rho_{2,3} = \frac{\sum_{y=2,3}(r_{2,y} - \bar{r}_2')(r_{3,y} - \bar{r}_3')}{\sqrt{\sum_{y=2,3}(r_{2,y} - \bar{r}_2')^2}\sqrt{\sum_{y=2,3}(r_{3,y} - \bar{r}_3')^2}}$$

$$= \frac{(1-2)(1-2) + (3-2)(3-2)}{\sqrt{(1-2)^2 + (3-2)^2}\sqrt{(1-2)^2 + (3-2)^2}}$$

$$= 1$$

同様に計算し，2：源田と 4 今井の相関係数は$\rho_{2,3} = -1$となる．

次に式(2)の推定評価値を計算する．まず，2：源田の全評価済みアイテムの平均評価値を求める．

$$\bar{r}_2 = (\sum_{y=2,3} r_{2,y})/2 = (1+3)/2 = 2$$

最後に，これまで計算してきた値を式(2)に代入する．

$$\hat{r}_{2,1} = \bar{r}_2 + \frac{\sum_{x=1,3,4}\rho_{2,x}(r_{x,1} - \bar{r}_x')}{\sum_{x=1,3,4}|\rho_{2,x}|}$$

$$= \frac{2 + 0(1-3) + 1(2-2) + (-1)(1-5/2)}{|0| + |1| + |-1|}$$

$$= 2.75$$

よってこの値から，2：源田の 1：とんかつへの推定評価値は 2.75 と計算することができ，この値は最大評価値 3 に限りなく近いため，2：源田はとんかつが好きであると予想できる．

③ 利用者間型メモリベース協調フィルタリングの問題点

前節までで，利用者間型メモリベース協調フィルタリングの手法と，その具体例について述べた．特に GroupLens の手法を用いることによって，与えられたレビューデータに対して予測を立て，ユーザーに対して推定評価値を提示することが可能となる．しかし，このモデルは対象となるユーザーの履歴が溜まっていない場合には対応しにくい，外れ値に影響されやすいといった問題点を抱えている．

そこで，本研究では，機械学習を用いた情報推薦システムのアルゴリズムを提案することで，数値予測の精度向上を図る．また，複数の評価指標を用いた比較を行うことによって研究の優位性の検証を行う．

第 4 章 使用データと前提条件

本章では，本論文で用いる使用データと前提条件について述べる．

4.1 実験データの概要

本節では，本研究で扱う実験データについて述べる．アルゴリズムを実装し，精度評価を行うためのデータとして， MovieLens[4]と，MyAnimeList 社が提供する Anime Dataset With Reviews[5]（以下 ADWR）を用いる．これらは，推薦システムと呼ばれるレビューデータの数値予測システムの開発やベンチマークのために作られた，レビューのためのデータセットである．MovieLens と ADWR でユーザができるアクションは基本的に以下の 3 つのみである．

- アイテムに評価を付ける
- アイテムをウィッシュリストに入れる
- アイテムにタグをつける

本研究では，それぞれのデータを加工し，全体の 70 パーセントを訓練用，残りの 30 パーセントをテスト用へと無作為に分割する．また得られた予測値をもとにアルゴリズムの精度評価を行い，それぞれの結果を比較することで研究の優位性の検証を行う．

4.2 実験データの考察

本節では，2 つの実験データの統計量や訓練データ，テストデータへの分割方法について述べる．

① 映画レビューデータ

本節では，本研究で映画レビューデータとして採用する MovieLens について述べる．MovieLens は Group Lens 社が学術目的としての利用に対し無償提供を行う映画鑑賞に対するレビューのデータである．

まず本研究では，MovieLens に格納されているデータのうち，映画を鑑賞した人物のタグである userId，視聴した映画のタグである movieId，視聴した映画を 0.5 から 5.0 の 0.5 刻みで評価したレビュー値である ratings の 3 つの列のみを取り出して用いる．取り出したデータの上位 3 件を表 2 に示す．

表 2 加工した映画レビューデータの上位 3 件

userId	movieId	ratings
1	1	4
1	6	4
1	50	5

また，取り出したデータの統計量は以下のようになっている．

表 3 映画レビューデータの統計量

データ数	100,836
userId の件数	962
movieId の件数	9,724

次に，取り出した映画データのうち，70 パーセントを訓練データに用いるため，train.movie という新たなデータセットとして格納する．train.movie の ratings タグのヒストグラムを以下の図 2 に示す．

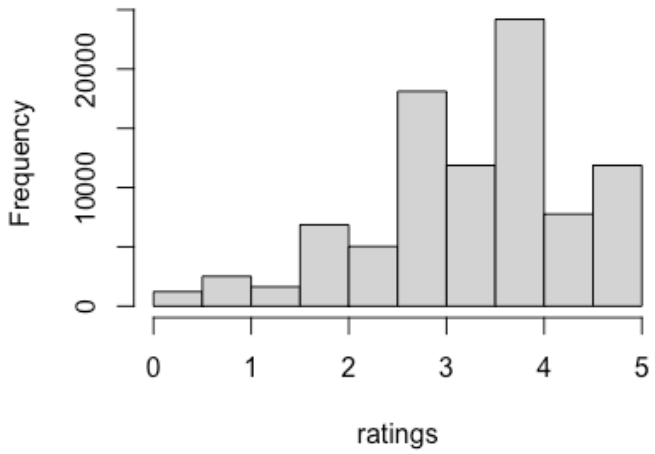

図 2 train.movie の ratings 分布

最後に，残りの 30 パーセントをテストデータとして test.movie に格納する．test.movie の ratings タグのヒストグラムを以下の図 3 に示す．

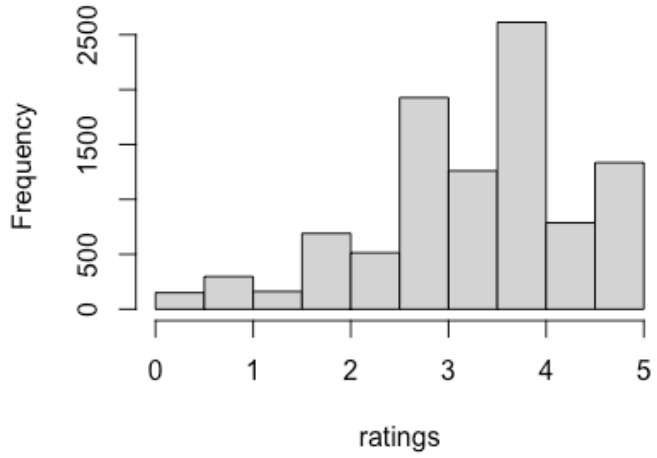

図 3 test.movie の ratings 分布

図 2 と図 3 を比較すると，訓練データ・テストデータのどちらも同じような分布になることが確認できる．また，図 2 と図 3 から，ratings が 3,4,5 となるデータが多くを占めていることが分かる．

② アニメレビューデータ

本節では，本研究でアニメレビューデータとして採用する ADWR について述べる．

ADWR は MyAnimeList 社が無償提供を行うアニメの視聴に対するレビューのデータである．

まず本研究では，ADWR に格納されているデータのうち，アニメを鑑賞した人物のタグである userId，視聴したアニメのタグである animeId，視聴したアニメを 1 から 10 の 1 刻みで評価したレビュー値である ratings の 3 つの列のみを取り出して用いる．取り出したデータの上位 3 件を以下に示す．

また，取り出したデータの統計量は以下のようになっている．

表 4 アニメレビューデータの統計量

データ数	192,110
userId の件数	130,517
movieId の件数	8,113

次に，取り出した映画データのうち，70 パーセントを訓練データに用いるため，train.anime という新たなデータセットとして格納する．train.anime の ratings タグのヒストグラムを以下の図 4 に示す．

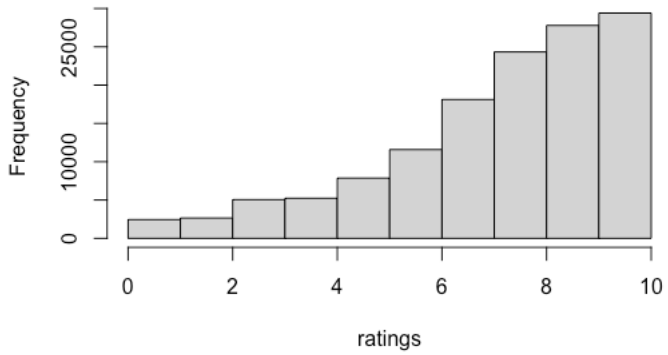

図 4 train.anime の ratings 分布

最後に，残りの 30 パーセントをテストデータとして test.anime に格納する．test.anime の ratings タグのヒストグラムを以下の図 5 に示す．

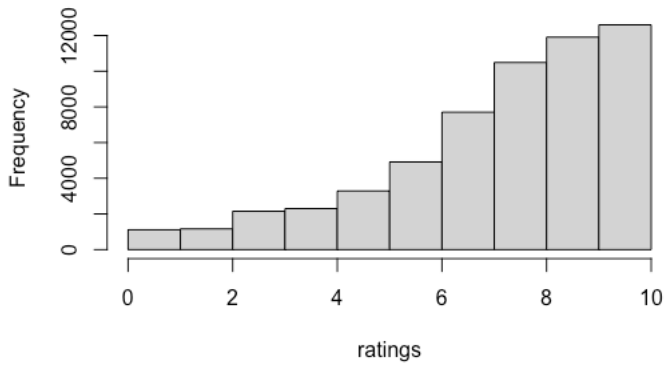

図 5 tests.anime の ratings 分布

4.3 実験環境

本節では，実験に用いたコンピュータの環境や，使用したライブラリに関して述べていく．

本研究のコンピュータの実験環境を表 5 に示す．

次に，本研究の情報推薦システムの構築に用いるライブラリについて説明する．本研究では，機械学習を用いた情報推薦システムの構築のために，既存の機械学習のライブラリである recosystem を利用した．recosystem は，R で記述された情報推薦システム構築の為のライブラリであり，本研究で利用したアルゴリズムである確率的勾配法が実装されている．

表 5 実験環境

OS	macOS Big Sur Ver.11.6
メモリ	16.0GB
Core	8
プログラム言語	R Ver. 4.1.1
機械学習ライブラリ	recosystem

第 5 章 実験内容

本章では，本研究における実験の方法について述べる．本研究では，提案するアルゴリズムの精度を高めるためにパラメータの最適化を行う実験と，最適化されたパラメータをもとにアルゴリズムの実装を行い，他のアルゴリズムとの精度比較を行う 2 つの実験工程を行う．

5.1 提案手法の精度向上

本節では，提案するアルゴリズムの精度を高めるための実験工程について述べる．機械学習では，パラメータの設定がアルゴリズムの精度を左右する．しかしながら，パラメータの決め方は確立されていないこと，タスクごとにパラメータが異なることから，実験によってパラメータを調整する必要がある．

本研究で決めるパラメータは以下のとおりである．

① 行列P, Qの次元k

図 5 のように行列分解によって得られた行列P, Qの次元kの値を決める．

100,200,300,400 と変更する．

② 正規化項のパラメータλ_P, λ_Q

目的関数に含まれる，過学習を防ぐための正規化項のパラメータλ_P, λ_Qの値を決める．

それぞれ 0.1,0.2,0.3,0.4 と変更する．

③ エポック数

データセットに含まれるデータが少なくとも 1 回は学習に含まれるように必要となる学習の回数である．10,50,100,150 と変更する．

5.2 実験で用いる数値予測モデル

本節では，映画・アニメレビューデータの数値予測モデルとして，4 つのモデルを紹介する．

① ランダムモデル

ランダムモデルは，予測値を評価値の範囲内で全て乱数値として与えたモデルである．ランダムモデルのステップは以下のようになる．

Step1. 訓練用データを乱数値として入力する．

Step2. テストデータとして同様に乱数値を与える．

② 平均値モデル

平均値モデルは，予測値をすべてのデータの平均値として与えたモデルである．平均値モデルのステップは以下のようになる．

Step1. 訓練用データの平均値を取得する
Step2. 得られた平均値をテスト用データとして入力する

③ 利用者間型メモリベース協調フィルタリングモデル

利用者間型メモリベース協調フィルタリングモデルは 3.2.1 節と 3.2.2 節で記述したとおりとなる．

④ 提案するモデル

本節では，本論文で提案する機械学習を用いたモデルについて提案する．まず，本モデルに用いる変数，定数を定義する．

k, λ_P, λ_Q: パラメータ
α: 学習率
m: ユーザ数
n: アイテム数
P: 潜在行列
Q: 潜在行列
$\hat{R}$: 予測評価点行列
$\boldsymbol{p}_i \in \mathbb{R}^k$ ： ユーザーの特徴行列
$\boldsymbol{q}_u \in \mathbb{R}^k$ ： アイテムの特徴行列

本モデルは，Matrix Factorization の手法を用いて予測評価点行列$\hat{R}$をユーザーの特徴行列$P \in \mathbb{R}^{k \times m}$とアイテムの特徴行列$Q \in \mathbb{R}^{k \times n}$に行列分解し，実測値と予測値の誤差が最小となる目的関数を解く．$\boldsymbol{p}_u \in \mathbb{R}^k$，$\boldsymbol{q}_u \in \mathbb{R}^k$となる場合において，$r_{u,i} \simeq \boldsymbol{p}_u^T \boldsymbol{q}_i$は行列$P$の$u$列目，行列$Q$の$i$列目を表す．このとき，目的関数は以下のようになる．

$$\text{Minimize} \sum_{(u,i) \in R} (r_{u,i} - \boldsymbol{p}_u^T \boldsymbol{q}_i)^2 + \lambda_P \|\boldsymbol{p}_u\|_F^2 + \lambda_Q \|\boldsymbol{p}_i\|_F^2 \tag{3}$$

この式は実測値と予測値の誤差に対して過学習を防ぐための正規化項を加えたものである．式(3)の最適化問題を解くためのアルゴリズムとして，本研究では確率的勾配降下法(stochastic gradient descent, 以下 SGD)[6]を採用する．SGD では，式(3)の勾配を全てのデータに対して調べるのではなく，(u, i)の候補をランダムに 1 つ取り出して誤差を計算し，パラメータの更新を行う．取り出した$r_{u,i}$によって，目的関数は以下の式によって更新される．

$$\left(r_{u,i} - \boldsymbol{p}_u^T \boldsymbol{q}_i\right)^2 + \lambda_P \boldsymbol{p}_u^T \boldsymbol{q}_i + \lambda_Q \boldsymbol{p}_i^T \boldsymbol{q}_u \tag{4}$$

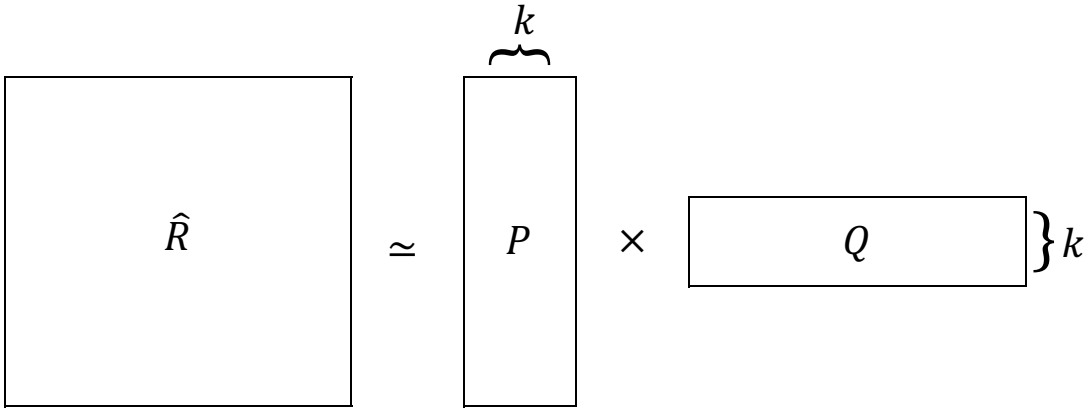

図 6 Matrix Factorization の概略図

$\boldsymbol{q}_u$と$\boldsymbol{q}_i$の勾配を計算した後，各変数は以下のようにして更新される．

$$\boldsymbol{p}_u \leftarrow \boldsymbol{p}_u + \alpha\left(e_{u,i}\boldsymbol{q}_i - \lambda_P \boldsymbol{p}_u\right),$$
$$\boldsymbol{q}_i \leftarrow \boldsymbol{q}_i + \alpha\left(e_{u,i}\boldsymbol{p}_u - \lambda_Q \boldsymbol{q}_i\right),$$

where

$$e_{u,i} = r_{u,i} - \boldsymbol{p}_u^T \boldsymbol{q}_i \tag{5}$$

以上の手法を機械学習によって実現する．訓練用モデルを学習させ，テスト用データで数値予測を行いながら予測精度を高めていく．

5.3 モデルの精度評価

モデルの精度評価をすることで，モデルの優位性を確認することができる．本節では，5.1 節のモデルの精度評価に用いる評価関数を示す．

① Root Mean Squared Error (RMSE)

RMSE を求める式は以下のようにして表される．

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{r_j} - r_j\right)^2}{N}} \tag{6}$$

r_j; ユーザjの実測値
$\hat{r_j}$: ユーザjの予測値
N: データ数

② Mean Absolute Error (MAE)

MAE を求める式は以下のようにして表される．

$$MAE = \frac{\sum_{i=1}^{N} |\hat{r_j} - r_j|}{N} \tag{7}$$

なお，与えられる定数と変数は 5.3.1 と同様である．

5.4 実験工程

本節では，提案するアルゴリズムとその他 3 つのアルゴリズムとの数値予測精度の比較の実験工程について述べる．本実験の工程は以下の順序で進める．

Step1. 2 つの実験データ（MovieLens，ADWR）の欠損値や外れ値を確認し，データの整形を行う．
Step2. 映画データのみを用いて提案手法のパラメータの最適値を算出する．
Step3. 2 つの実験データを用いて，Step2 でパラメータの最適化を行った提案手法のモデルを含む 4 つのモデルの実装を行い，レビュー値の予測を行う．
Step4. テストデータを用いて，Step3 で実装したモデルの数値予測精度をそれぞれ算出する．

第 6 章 実験結果

本章では，5 章に記述した内容をもとに実験を行った結果を記述する．

6.1 予測精度向上のためのパラメータ調整

本節では，実験工程の 1 つであるパラメーター調整の実験結果を掲載する．各パラメータの初期値を表 6 のように設定し，左から順に 1 つずつパラメータの分類精度が高くなるように調整していく．また，本節の実験で用いるデータは映画レビューデータのみとする．

初期設定をもとに各パラメーターの最適化を行った結果は以下のとおりである．

表 6 パラメータの初期設定

	k	λ_P	λ_Q	イテレーション
値	100	0.2	0.2	50

表 7 行列P, Qの次元k

k	RMSE	MAE
100	0.901	0.699
200	0.887	0.686
300	0.900	0.693
400	0.892	0.686

結果からk =200 の場合において最も高い予測精度が出ることが確認できる．よってこの値を採用する．

表 8 正規化項のパラメータλ_P

λ_P	RMSE	MAE
0.1	0.851	0.658
0.2	0.859	0.667
0.3	0.874	0.682
0.4	0.888	0.695

表 9 正規化項のパラメータλ_Q

λ_Q	RMSE	MAE
0.1	0.849	0.656
0.2	0.856	0.665
0.3	0.866	0.675
0.4	0.879	0.688

結果からλ_P, λ_Q=0.1 の場合において最も高い予測精度が出ることが確認できる．

表 10 エポック数

エポック数	RMSE	MAE
10	0.874	0.675
50	0.851	0.658
100	0.853	0.661
150	0.855	0.662

結果から，エポック数が 50 のとき最も高い予測精度が出ることが確認できる．

以上の結果をもとに算出された各パラメータの最適値は以下のとおりである．

表 11 パラメータの最適値

	k	λ_P	λ_Q	イテレーション
値	200	0.1	0.1	50

6.2 数値予測精度の比較

本節では，6.1 節で得た提案モデルのパラメータの最適値によって算出された数値予測精度と，その他のモデルの数値予測精度の比較を行う．評価関数には 5.3 節で述べた RMSE，MAE を用いる．映画・アニメレビューデータの数値予測の実験結果はそれぞれ以下のとおりである．

表 12 映画レビューデータの数値予測の実験結果

アルゴリズム	RMSE	MAE
ランダム	1.932	1.575
平均値	1.054	0.834
協調フィルタリング	1.072	0.829
提案するモデル	0.849	0.656

表 13 アニメレビューデータの数値予測の実験結果

アルゴリズム	RMSE	MAE
ランダム	4.218	3.427
平均値	2.264	1.813
協調フィルタリング	2.036	1.535
提案するモデル	1.830	1.258

6.3 数値予測精度に対する考察

本節では，6.2 節で行った数値予測精度の比較に関する考察を行う．表 6，表 7 の結果から，映画・アニメレビューデータのどちらにおいても，提案手法の予測精度が最も優れていることが分かる．また，RMSE，MAE のどちらの評価関数においても本研究の提案するモデルの優位性を確認することができた．

一方で，2 つのデータにおいて予測精度に大きな差が生じていることが分かる．映画レビューデータは RMSE，MSE ともに 1.0 を下回っている一方で，アニメレビューデータにおいてはどちらも 1.0 を超えた数値となっている．この差が生じた原因として考えられるのが 2 つのデータが持つ特徴の違いである．表 3 と表 4 を比較すると，映画レビューデータはアイテムをレビューしたユーザー1 人あたりの件数が約 105 件であるのに対し，アニメレビューデータは約 1.5 件にとどまっている．

ユーザー1 人あたりのレビュー件数が予測精度に影響を与えるかどうかを検証するため，映画レビューデータを加工し，レビュー件数が多いユーザー上位 25%，50%のデータを削除し，再度実験を行った結果を以下に示す．

表 14 加工した映画レビューデータの実験結果

データ	RMSE	MAE
元データ	0.849	0.656
レビュー件数上位 25%のユーザーを削除したデータ	1.055	0.821
レビュー件数上位 50%のユーザーを削除したデータ	1.082	0.841

実験の結果，レビュー件数が多いユーザーを削除するにつれて予測精度が落ちていることが分かる．このことから，本研究の提案手法の精度を高めるためには，ユーザー1 人あたりが多くのアイテムのレビューを行ったデータを扱うことが 1 つの有効策であるということがわかる．

第 7 章 おわりに

本研究では，インターネットによるサービス需要の増加とともに，ユーザーのレビューデータがサービスの供給者に与える影響力が高まるという予想を踏まえ，機械学習を用いた数値予測モデルを提案した．そして，インターネットサービスを提供する供給者がユーザーに最適なアイテムの推薦を行うための一つの解決策となることを目的として研究を行なった．

第 2 章では，本研究の提案手法である機械学習の概要と種類について述べた．

第 3 章では，本研究の関連知識と先行事例について述べた．先行事例として利用者間型メモリベース協調フィルタリングを紹介し，このモデルの持つ特徴と問題点を挙げた．

第 4 章では，本研究で用いた 2 つの実験データである映画レビューデータ・アニメレビューデータを紹介し，各データの統計量と特徴について述べた．また，本研究の実験環境についても触れた．

第 5 章では，本研究の数値予測実験に用いるモデルの紹介と予測の精度評価方法について述べた．提案手法では，行列分解を行うためのアルゴリズムとして確率的勾配降下法を採用し，機械学習を用いてモデルの実装を行った．

第 6 章では，パラメータの最適化を行った提案手法と，その他 3 つのモデルを用いて，レビュー値の数値予測実験を行った．数値予測を行う対象として，映画とアニメレビューデータの 2 種類を扱い，どちらのデータの適応においても提案手法の優位性が確認できた．また，2 つのデータの予測精度を比較すると，映画レビューデータにおいて高い予測精度が得られた．これは，ユーザー1 人あたりがどれだけ多くのアイテムに対してレビューを行ったかが関係していることが確認できた．

今後，更にモデルの改良を重ね，予測精度を向上させることがユーザーの満足度を高める情報推薦システム構築の実現に繋がるだろう．また，本研究では約 10〜20 万件のデータを扱ったが，より膨大なデータを扱う際には計算速度にも着目する必要があると考えられる．

本研究は，今後レビュー値をもとにユーザーに対して最適なサービスを提供するための 1 つの解決策となると考えられる．また数値予測を行う前段階であるデータの収集において考慮するべき点についても，本研究を役立てることができるだろう．

参考文献

[1] 財団法人インターネット協会，“インターネット白書 2012”，インプレス R&D，p.6，2013

[2] 松本実大，黒木啓之，“Deep Learning における過学習の及ぼす影響の評価”，第 21 回電子情報通信学会東京支部学生会研究発表会論文集，p.148，2016

[3] 神嶌敏弘，“推薦システムのアルゴリズム”，pp.53-57，2016
（参照：2021 年 12 月 15 日）
https://www.kamishima.net/archive/recsysdoc.pdf

[4] Grouplens，“MovieLens Latest Datasets”
（参照：2021 年 9 月 28 日）
https://grouplens.org/datasets/movielens/latest/

[5] MyAnimeList，“Anime Dataset with Reviews”
（参照：2021 年 11 月 15 日）
https://www.kaggle.com/marlesson/myanimelist-dataset-animes-profiles-reviews?select=reviews.csv

[6] Zhuang，W-S Chin，Y-C Juan，and C-J Lin，“A fast parallel SGD for matrix factorization in shared memory systems”，Publication:RecSys ‘13: Proceedings of the 7th ACM Conference on Recommender Systems，pp.249-256，2013.