lp 正則化項を用いた 3 次元テンソル分解

プロフィットデザイン研究　　5215F007-6　神田大成
指導教員　大野髙裕

3-dimensional tensor factorization using lp regularizer

KANDA Taisei

1. 研究背景と目的

近年, ビッグデータをマーケティングに活かす動きに注目が集められている. 例えば, TSUTAYA は T ポイントカードに蓄積された顧客情報を使用し, T ポイント加盟店を訪れた顧客に対して, レシートクーポンを配布することで販売促進を行っている. 他の事例として, Amazon は, 顧客の行動データを蓄積し, おすすめの商品を掲示することで販売促進を行っている.

上記の事例のように, 顧客にお勧めの商材を掲示し, 次の購買行動を喚起するものとして推薦システムがある. 推薦システムに関する手法の一つとして, 協調フィルタリングという手法がある. これは, 対象者の商材購買データと対象者以外の商材購買データ両方を用い, 顧客の購買パターンからユーザ間, 又は商品間の類似性を解析しパーソナライズな推薦を行なう手法である[1].

koren et al.[2] は, 行列因子分解の手法を推薦システムに適用することで, 顧客データと商品データで構成される 2 次元行列を顧客, 商品, それぞれの 2 次元行列に分解し, 2 つの特徴量を捉えた 2 次元における協調フィルタリングを可能にしている. 近年, より高度な推薦を行うため, 顧客・商品情報に加えて場所や時間等の文脈情報を考慮した 3 次元以上の推薦システムへの関心が高まってきている. テンソルを用いて行列の次元を拡大し, それぞれの特徴量を持つ 2 次元行列に分解することにより, 2 つ以上の要因を考慮した推薦システム構築が可能になった.

一方で, 実際に使用するデータは疎なデータであることが多いという問題がある. 捉えることのできる特徴量が増加し, データの次元数が増えていくと, 変数の数は増加する一方で, データの疎性は増加していき, 精度の良い推定を行なうことが難しくなる. 効果的な推薦を行なうためには, 膨大なデータの中から意味のある変数のみを選択していくことが重要な課題である. Karatzoglou et al.[3] は, 不要な説明変数の推定量を縮小させる働きを持つ正則化項を導入することで, 疎なデータに対応した. しかし, Karatzoglou et al. が用いている正則化項モデルは, 不要変数の推定量を限りなくゼロに縮小させる働きを持つが, 完全にゼロにする働きを持たない. つまり, 変数選択まで完全に行うことができないという課題がある.

本研究では, Karatzoglou et al. と同様の 3 次元テンソル分解を用いた推薦システムのモデルに, 不要な説明変数の推定量をゼロにする働きを持つ正則化項を導入したモデルを構築する. そして, 従来研究で提案されている正則化項モデルと推定精度の差を比較し, 本研究で提案した正則化項モデルの優位性を示すことを目的とする

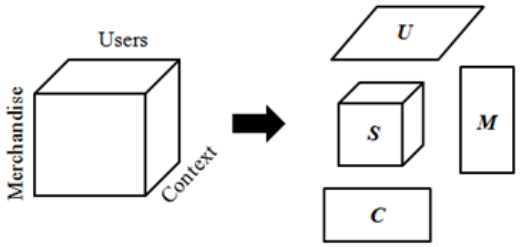

図 1.higher-order singular value decomposition[4]

2. 従来研究

2.1. HOSVD

テンソル分解のモデルには複数の種類が存在する. 代表的なものとして CP 分解や HOSVD[4] が挙げられる. HOSVD 分解は, 図 1 のように 3 次元行列を, $\boldsymbol{U} \in \mathbb{R}^{u \times d_U}$, $\boldsymbol{M} \in \mathbb{R}^{M \times d_M}$, $\boldsymbol{C} \in \mathbb{R}^{C \times d_C}$ の 3 つの行列と, 中央テンソル $\boldsymbol{S} \in \mathbb{R}^{d_U \times d_M \times d_C}$ に分解する. ここで, $\boldsymbol{U}$, $\boldsymbol{M}$, $\boldsymbol{C}$ はそれぞれ, 顧客, 商品, 場所や購買時間等の文脈情報, それぞれの特徴量を持った行列を表す. 添え字 u, m, c は各行列の行数を, d_U, d_M, d_C は各行列の列数を表す. HOSVD は以下のように定式化される.

$$F_{ijk} = \boldsymbol{S} \times_U \boldsymbol{U_{i*}} \times_M \boldsymbol{M_{j*}} \times_C \boldsymbol{C_{k*}}. \quad (1)$$

ここで, $\times_U$, $\times_M$, $\times_C$ は, 各テンソルの向きを表し, $\boldsymbol{U_{i*}}$, $\boldsymbol{M_{j*}}$, $\boldsymbol{C_{k*}}$ は, 各行列の行を示している. この分解手法は, d_U, d_M, d_C のパラメータを調整することで, 顧客, 商品, 文脈情報行列のランク数を調整することができる. これにより、使用するデータセットが大きくなってもランク数を調整することで、分解する行列の推定が可能になる.

2.2. Karatzoglou et al.

HOSVD には問題が 2 点ある. 1 点目は, データが密であるという仮定の元に計算を行っているため, データの疎性を考慮していないことである. 上記の仮定の元で計算する場合, 評価されていないデータをすべてゼロとして処理する必要がある. しかし, 評価した値がゼロであることと, 評価されていないデータをゼロとすることを同等に扱うべきではない. この様な処理を行うと, 推定に偏りが生じる可能性がある. 2 点目は, データの次元数が増えるほど, 推定する変数の数が多くなりモデルが複雑化するということである. 上記の場合, サンプルデータに対してはよくあてはまるモデルが構築されるが, 未知のデータに対する予測精度が悪化する過学習が生じる場合がある.

Karatzoglou et al. は, HOSVD のモデルに対し, 目的関数に不要な説明変数の推定量を縮小させる働きを持つ正則化項を組み込んだモデルを構築し, HOSVD の問題に対応している.

Karatzoglou et al. では損失関数，正則化項の和で構成される目的関数の最小化を行なっている；

$$R[\boldsymbol{U},\boldsymbol{M},\boldsymbol{C},\boldsymbol{S}] \coloneqq L(F,Y) + \Omega[\boldsymbol{U},\boldsymbol{M},\boldsymbol{C}] + \Omega[\boldsymbol{S}]. \quad (2)$$

$L(F,Y)$は損失関数，$\Omega[\boldsymbol{U},\boldsymbol{M},\boldsymbol{C}]$，$\Omega[\boldsymbol{S}]$は正則化項を表す．ここでは元データの評価値と，推定した評価値の差分を損失関数として定めている；

$$L(F,Y) := \frac{1}{\|S\|_1}\sum D_{ijk} L\left(F_{ijk,} Y_{ijk}\right). \quad (3)$$

ここで，F，Yはそれぞれ推定した評価値，元データの評価値を表す．Fについては，(1)式で求めることができる．Sは，Fig.1 における中央テンソル，D_{ijk}は非負数であり，Y_{ijk}が観測される時のみ 1 を与える二値変数である．D_{ijk}を与えることにより，観測されたデータのみの最適化を行なうことができる．この二値変数を与えることで，前述した HOSVD の 1 点目の問題である，データの疎性に対応している．$L(F_{ijk,}Y_{ijk})$ は，元データと推定値の差分を表す二乗誤差関数であり，以下の式で表される；

$$L\left(F_{ijk,}Y_{ijk}\right) = \frac{1}{2}\left(f_{ijk} - y_{ijk}\right)^2. \quad (4)$$

Karatzoglou et al.は、過学習を抑えるために，目的関数に対して正則化項を定めている；

$$\Omega[\boldsymbol{U},\boldsymbol{M},\boldsymbol{C}] \coloneqq \frac{1}{2}[\lambda\|\boldsymbol{U}\|^2 + \lambda\|\boldsymbol{M}\|^2 + \lambda\|\boldsymbol{C}\|^2], \quad (5)$$

$$\Omega[\boldsymbol{S}] \coloneqq \frac{1}{2}[\lambda\|\boldsymbol{S}\|^2]. \quad (6)$$

λは正則化パラメータであり，正則化項の強さを調整することができる．

2.3. 正則化項

Karatzoglou et al.の(5)，(6)式で用いられている正則化項は以下の式で表される．

$$\frac{1}{p}\lambda\|\boldsymbol{\beta}\|^p. \quad (7)$$

ここで，λは正則化パラメータ，$\boldsymbol{\beta}$は推定する変数を表し，pの値によって，正則化項の種類は異なる．

2.3.1. ridge 回帰

(7)式において，p=2 の場合を ridge 回帰(ℓ_2正則化項)[5]と呼び，これは Karatzoglou et al. が提案モデルに用いている正則化項と一致する．

ridge 回帰は，不要な説明変数の推定量を縮小させる働きを持つ。説明変数を x1，x2 の 2 つとした際の具体例を図 2 に示す．誤差項と正則化項の和を目的関数とした際に，図 2 の太線で結んだ部分が最小化の解となる．この際，x1，x2 それぞれの説明変数の推定量を縮小させることができている．ridge 回帰は正則化項部分を微分することで，解析的に解くことができるため，計算は比較的簡単なものになり適用しやすい．しかし，ridge 回帰では不要な変数を限りなくゼロに縮小させることはできるが，完全にゼロにする働きを持たない．

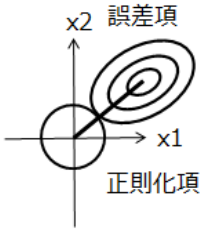

図 2.ridge 回帰

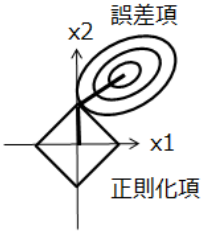

図 3.lasso

2.3.2. lasso

ridge 回帰の他に代表的なものとして lasso(ℓ_1正則化項)[6] がある．これは，(7)式において，p=1 とした際の正則化項である．説明変数を x1，x2 の 2 つとした際の具体例を図 3 に示す．誤差項と正則化項の和を目的関数とした際に，図 3 の太線で結んだ部分が最小解となる．この際，x1 の推定量を完全にゼロにしている．

このようにlassoを用いることで，いくつかの説明変数の推定量を完全にゼロにすることができる．不要な説明変数を完全にゼロにすることで，モデルの複雑性が緩和され，過学習を防ぐことができるだけでなく，高速計算を可能にする．しかし，lassoは評価関数に絶対値を含むため解析的に解くことができないため，計算の際に式変形が必要である．

3. 提案モデル

3.1. モデリング

本研究では，karatzoglou et al. で提案されたモデルに対して，2.3 節で述べた lasso を組み込んだモデルを提案する．テンソルの分解式，損失関数，目的関数は，前述した，(1)，(2)，(3)，(4)式を用いる．正則化項部分のモデル式は以下のようになる；

$$\Omega[\boldsymbol{U},\boldsymbol{M},\boldsymbol{C}] \coloneqq [\lambda\|\boldsymbol{U}\| + \lambda\|\boldsymbol{M}\| + \lambda\|\boldsymbol{C}\|], \quad (8)$$

$$\Omega[\boldsymbol{S}] \coloneqq [\lambda\|\boldsymbol{S}\|]. \quad (9)$$

正則化項部分は上記のように絶対値を取る．

3.2. 推定方法

推定には確率的勾配降下法を用いる．これはデータセットからランダムにサンプリングしたデータの勾配を計算し，パラメータの更新を行う手法である．毎回データ全体の勾配を計算してパラメータ更新を行う勾配降下法と違って，サンプリングしたデータ単位で勾配を計算して更新を行うため，メモリ効率が良い．

パラメータの更新に必要な各行列$\boldsymbol{U}$，$\boldsymbol{M}$，$\boldsymbol{C}$，$\boldsymbol{S}$の勾配を，損失関数$L(F,Y)$を$\boldsymbol{U}_{i*}$，$\boldsymbol{M}_{j*}$，$\boldsymbol{C}_{k*}$，$\boldsymbol{S}$それぞれについて

偏微分することで求める；

$$\partial_{U_{i*}} L(F_{ijk}, Y_{ijk}) = \partial_{F_{ijk}} L(F_{ijk}, Y_{ijk}) S \times_M M_{j*} \times_C C_{k*}, \quad (10)$$

$$\partial_{M_{j*}} L(F_{ijk}, Y_{ijk}) = \partial_{F_{ijk}} L(F_{ijk}, Y_{ijk}) S \times_U U_{i*} \times_C C_{k*}, \quad (11)$$

$$\partial_{C_{k*}} L(F_{ijk}, Y_{ijk}) = \partial_{F_{ijk}} L(F_{ijk}, Y_{ijk}) S \times_U U_{i*} \times_M M_{j*}, \quad (12)$$

$$\partial_S L(F_{ijk}, Y_{ijk}) = \partial_{F_{ijk}} L(F_{ijk}, Y_{ijk}) U_{i*} \otimes M_{j*} \otimes C_{k*}. \quad (13)$$

⊗はテンソルの外積を示す．(10), (11), (12), (13)式で求めた勾配を元に，各パラメータの更新式を以下に定める；

$$U_i^{t+1} = \begin{cases} U_i^t - \eta_t \partial_{U_{i*}} L(F_{ijk}, Y_{ijk}) - \lambda\eta_t, \\ where\ U_i^t - \eta_t \partial_{U_{i*}} L(F_{ijk}, Y_{ijk}) > \lambda\eta_t, \\ U_i^t - \eta_t \partial_{U_{i*}} L(F_{ijk}, Y_{ijk}) + \lambda\eta_t, \\ where\ U_i^t - \eta_t \partial_{U_{i*}} L(F_{ijk}, Y_{ijk}) \le \lambda\eta_t, \\ 0, \\ where\ otherwise, \end{cases} \quad (14)$$

$$M_j^{t+1} = \begin{cases} M_j^t - \eta_t \partial_{M_{j*}} L(F_{ijk}, Y_{ijk}) - \lambda\eta_t, \\ where\ M_j^t - \eta_t \partial_{M_{j*}} L(F_{ijk}, Y_{ijk}) > \lambda\eta_t, \\ M_j^t - \eta_t \partial_{M_{j*}} L(F_{ijk}, Y_{ijk}) + \lambda\eta_t, \\ where\ M_j^t - \eta_t \partial_{M_{j*}} L(F_{ijk}, Y_{ijk}) \le \lambda\eta_t, \\ 0, \\ where\ otherwise, \end{cases} \quad (15)$$

$$C_k^{t+1} = \begin{cases} C_k^t - \eta_t \partial_{C_{k*}} L(F_{ijk}, Y_{ijk}) - \lambda\eta_t, \\ where\ C_k^t - \eta_t \partial_{C_{k*}} L(F_{ijk}, Y_{ijk}) > \lambda\eta_t, \\ C_k^t - \eta_t \partial_{C_{k*}} L(F_{ijk}, Y_{ijk}) + \lambda\eta_t, \\ where\ C_k^t - \eta_t \partial_{C_{k*}} L(F_{ijk}, Y_{ijk}) \le \lambda\eta_t, \\ 0, \\ where\ otherwise, \end{cases} \quad (16)$$

$$S^{t+1} = \begin{cases} S^t - \eta_t \partial_S L(F_{ijk}, Y_{ijk}) - \lambda\eta_t, \\ where\ S^t - \eta_t \partial_S L(F_{ijk}, Y_{ijk}) > \lambda\eta_t, \\ S^t - \eta_t \partial_S L(F_{ijk}, Y_{ijk}) + \lambda\eta_t, \\ where\ S^t - \eta_t \partial_S L(F_{ijk}, Y_{ijk}) \le \lambda\eta_t, \\ 0, \\ where\ otherwise, \end{cases} \quad (17)$$

$$\eta_t = \frac{\eta_0}{\sqrt{t}}. \quad (18)$$

表 1.使用するデータセット

User	Movie	Time	Scale	Ratings
84	192	3	1〜13	1464

λは正則化パラメータを表し，ηは更新のステップ幅を決定するハイパーパラメータで一般に学習率と呼ばれている．学習率の初期値η_0は 0.01 とする．tは何回目にサンプリングしたデータかを表す．

本研究提案モデルでは，lasso の計算を，Duchi et al.[7]の手法を用いて絶対値の大小で場合分けをして計算を行う．これにより，$|\lambda\eta_t|$より値が小さくなった変数をゼロに縮小させることができる．

4. 検証

4.1. 使用データ

Karatzoglou et al. が実験に使用した，大学生の映画評価データ 1464 件を用いる．これは，大学生 84 人が最近観た映画を 1-13(13 が最高評価)の範囲で点数付けしたものである．また，映画の評価を行う際に，いつ観た映画かを「平日」，「休日」，「覚えていない」の 3 択の中から選んでもらい文脈情報として記録している．データセットの概要を表 1 にまとめる．

本研究では，ユーザデータ 84 種類を$\boldsymbol{U}$として，映画データ 192 種類を$\boldsymbol{M}$として，時間データ 3 種類を$\boldsymbol{C}$として，特徴行列の推定を行う．

4.2. 検証方法

本研究では，確率的勾配降下法を用いて Karatzoglou et al. の ridge 回帰モデルと，本研究の lasso モデルのパラメータ推定を行い，精度の評価を RMSE(Root Means Square Error)を用いて行う．これは予測値が正解からどの程度乖離しているかを表す指標で，0 に近い値であるほど予測精度が優れている．計算式は以下に示す．

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \quad (19)$$

Nはデータ数，y_iは実測値，$\hat{y}_i$は予測値を表し，実測値と予測値の差の二乗和を計算している．

この際，5-fold Cross Validation を行うために使用データセットを 5 つに分割する．5 分割したしたデータセットのうち 4 つを推定用データセット，1 つを検証用データセットとして用い，5 分割したデータセットすべてが検証用データセットになるまで，合計 5 回検証を行い，5 回の結果を平均したものを RMSE の評価値とする．

また，推定する行列$\boldsymbol{U}$，$\boldsymbol{M}$，$\boldsymbol{C}$，$\boldsymbol{S}$のランク数を決定するパラメータを$d_U = d_M = d_C = d_S = 4$と定める．正則化パラメータ$\lambda$の値を 0.1 から 0.9 の範囲で動かすことで，正則化頂の強さと推定精度の変化を考察する．

4.3. 検証結果

RMSE による推定精度の比較結果を以下に示す．lasso, ridge 回帰，それぞれについてλの値を 0.1 から 0.9 の値まで調整し，RMSE の値が最も良かったもの，そしてλの変化による RMSE の推移を以下に示す．

表 2.RMSE による推定精度比較

	推定データ	検証データ	差分
ridge回帰	2.555	2.876	0.321
lasso	**2.593**	**2.803**	**0.21**

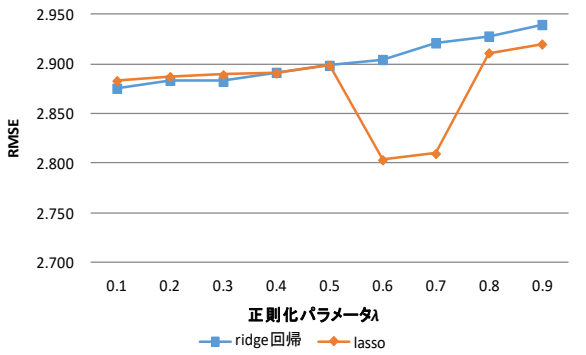

図 4.λの変化と RMSE の推移

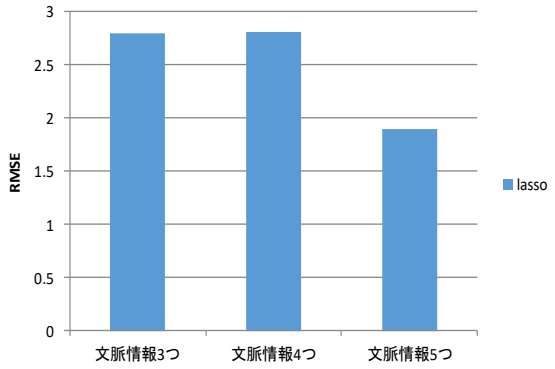

図 5.文脈情報数の変化と RMSE の推移

表 2 を見ると，従来の ridge 回帰と比較して，推定データ，検証データどちらにおいても RMSE の値が良いことがわかる．また，未知データに対する影響を確かめるために，推定データの RMSE と，検証データの RMSE の差分を取った．差分においても，本研究提案モデルの方が小さいことがわかる．この結果より，本研究で提案したモデルの，不要変数の推定量をゼロにする働きにより，モデルの複雑性を緩和することが出来たことがわかる．

図 4 は，λの値を 0.1 から 0.9 まで調整し，RMSE の変化の推移をプロットしたものである．図 4 の結果を見ると，λの値が小さいと，2 つのモデルの推定精度の差は小さくなることがわかる．推定精度は lasso の場合，$\lambda = 0.6$の時に最も良い値を取り，ridge 回帰の場合，$\lambda = 0.1$の時に最も良い値を取ることがわかる．また，risge 回帰，lasso，どちらも，λの値が最も小さい時が最も良い RMSE の値を取っている．

更なる考察を行うために，検証に使用する文脈情報を複数用意し，RMSE の差の比較を行った．その結果を図 5 に示す．一番左の「文脈情報 3 つ」は，本研究で最初から用いている時間データ，真ん中の「文脈情報 4 つ」は同伴者データであり，「一人」，「友達・彼女/彼女」，「家族・それ以外」，「覚えていない」の 4 種類ある．一番右の「文脈情報 5 つ」は季節データであり，「春」，「夏」，「秋」，「冬」，「覚えていない」の 5 種類ある．図 5 を見ると，「文脈情報 5 つ」を使用した際に，大きく推定精度が良くなっていることがわかる．この結果から，考慮する情報数が増えるほど，不要な変数の推定量をゼロにする正則化項モデルは効果を発揮することがわかる．

5. おわりに

本研究では，3 次元のテンソル分解を用いた推薦システムにおいて，不要な説明変数の推定量をゼロにする働きを持つ正則化項をモデルに導入することで，従来のモデルと比較して推定精度において優位性を示すことができた．

今後の課題は大きく 2 点ある．1 点目は，より大規模なデータになった際にどのような結果になるかを試す必要がある．今回使用したデータサイズは実務で使用するデータと比べると比較的小さなデータであり，データの疎性，変数の数，共に少ない．本研究で提案したスパース性に強いモデルの強みは，不要変数の除外によるモデルの簡略化と計算の高速化なので，大規模なデータの際はより効果を発揮すると考える．2 点目は，$\mathrm{lp}(0 < p \leq 1)$の条件下でのモデルの構築である．$p$=1 以下の正則化項は理論上，lasso より強いスパース性を持つため，更なる推定精度の向上が見込める．

上記 2 点の課題を踏まえた今後の展望は，異なる大きさのデータセットを複数用意し，ridge 回帰，lasso，$\mathrm{lp}(0 < p \leq 1)$条件下での正則化項それぞれを比較することで，データセットのサイズと適用すべき正則化項の関係性を分析する検証が考えられる．

参考文献

[1] 神嶌敏弘：“推薦システムのアルゴリズム，” 人口知能学会誌,Vol. 22, No.6, pp.826-837 (2007)

[2] Koren, Y., Bell, R. and Volinsky, C.: “Matrix Factorization Techniques for Recommender Systems,” *Computer*, Vol.42, No.8, pp.30-37 (2009)

[3] Karatzoglou, A., Amatriain, X., Baltrunas, L. and Oliver, N.: “Multiverse Recommendation : N-dimensional Tensor Factorization for Context-aware Collaborative Filtering,” *Proceedings of the Fourth ACM Conference on Recommender systems*, pp.79-86 (2010)

[4] Lathauwer,D, L., Moor,D, B. and Vandewalle, J : “A Multilinear Singular Value Decomposition,” *SIAM Journal on Matrix Analysis and Applications*, Vol.21, No.4, pp.1253-1278 (2000)

[5] Hoerl, A., Kennard, R.: “Ridge Regression: Biased Estimation for Nonorthogonal Problem,” *Technometrics*, Vol.12, pp.55-67 (1970)

[6] Tibshirani, R: “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society*, pp.267-288 (1996)

[7] Duchi,J., Singer,Y., : “Efficient online and batch learning using forward backward splitting,” *Journal of Machine Learning Research*, Vol.10, pp.2899–2934(2009)