自己符号化器の中間表現を用いた特徴分析に関する研究

1X14C033-2　金澤 真平
指導教員　後藤 正幸

1　研究背景と目的

近年，インターネット上で商品 (以下，アイテム) の売買を行う EC サイトを通じて，運営企業は膨大な購買履歴データを蓄積可能となった．また，EC サイト上の購買履歴データには，自社アイテムに関する情報の他に，利用者に関する情報などが付与されるようになっている．このような背景のもと，アイテムを類似性によってクラスタリングし，様々なマーケティング施策につなげる試みが行われている．

一方，本研究で対象とする EC サイト A は，様々な中古ファッションアイテムを扱っている．本サイトで扱われているアイテムは，非常に多種多様であり，様々な施策や適正な価格決定のためには，ある程度のクラスタリングが有効であると考えられる．しかし，各アイテムにはブランドや出品状態などといった属性や定性情報が付与されており，これらの多様な情報をそのままダミー変数化すると，各アイテムに対する特徴量が膨大になってしまう．このような高次元データに対して，一般的なクラスタリング手法を適用した場合，高次元空間において各データ間の距離がほぼ等しくなり，アイテム間の類似度をうまく表現できないという問題がある．

そこで，高次元データに対して予め次元圧縮を行ったもとで，クラスタリングすることを考える．一般的な次元圧縮手法として，線形主成分分析 (線形 PCA) やカーネル主成分分析 (カーネル PCA) が広く知られている．しかし，線形 PCA は，線形次元圧縮手法であり，本研究で対象とするデータのように非線形である場合，有効な次元圧縮が行えないという問題がある．また，代表的な非線形次元圧縮手法の一つであるカーネル PCA は，グラム行列の固有値計算などで計算コストがかかるため，実用が容易でないという問題がある．

そこで本研究では，高次元非線形データのクラスタリングを可能とするため，効率的な非線形次元圧縮を可能とする自己符号化器 (Autoencoder)[1] を用いたクラスタリング手法を提案する．また，実際の EC サイトにおける購買履歴データを用いた分析を行い，提案手法の有効性を示す．

2　準備

2.1　分析対象データ

本研究では，中古アイテムの買取りおよび再販売のビジネスを展開するファッション EC サイト A における購買履歴データを対象とし，次元圧縮手法を用いたクラスタリングを目的とする．各アイテムにはカテゴリ，ブランドのほか，出品する際の価格 (出品価格) や購入された際の価格 (販売価格)，割引率など様々な情報が付与されている．

表 1: 対象データの概要

変数	説明	ユニーク数	変数	説明	ユニーク数
性別	男性，女性	2	サイズ	例)S，M	589
大カテゴリ	例) アウター	6	色	例) 赤，白	90
小カテゴリ	例) ダウンコート	275	出品月	例)1 月，2 月	12
ブランド		20	出品価格		1
状態	例)1，2	5	販売価格		1
素材	例) ポリエステル	18,698	割引率		1

本サイトでは，アイテムの在庫削減のため，一定期間購入されなかった出品アイテムに対して値下げを行っており，どのようなアイテムがどの程度の価格で購入されるのかということに着目した分析を行うことは，機会損失を防ぐという観点において重要であると考えられる．そこで，本研究ではアイテムの割引率に着目した分析を行う．

2.2　自己符号化器

自己符号化器はニューラルネットワークのモデルを採用した次元圧縮手法の一種で，図 1 で表されるように入力層，中間層，出力層の 3 層構造からなるモデルである．

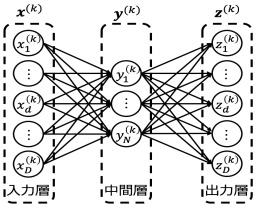

図 1: 自己符号化器

自己符号化器への K 個の入力データを $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(k)}, \ldots, \boldsymbol{x}^{(K)})^{\mathrm{T}} (\boldsymbol{x}^{(k)} \in \mathbb{R}^D)$ と表す．自己符号化器は，出力 $\boldsymbol{z} = (\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(k)}, \ldots, \boldsymbol{z}^{(K)})^{\mathrm{T}} (\boldsymbol{z}^{(k)} \in \mathbb{R}^D)$ が入力を再現することを目的としている．また，k 番目の入力データ $\boldsymbol{x}^{(k)}$ に対する N 次元の中間層ベクトルを $\boldsymbol{y}^{(k)} = (y_1^{(k)}, \ldots, y_n^{(k)}, \ldots, y_N^{(k)})^{\mathrm{T}}$ としたとき，n 番目のニューロン $y_n^{(k)}$ は式 (1) で表される．

$$y_n^{(k)} = f\left(\sum_{d=1}^{D} W_{dn} x_d^{(k)} + b_n\right) \tag{1}$$

ここで，$\boldsymbol{W} = \{W_{dn}\} \in \mathbb{R}^{D \times N}$，$\boldsymbol{b} = \{b_n\} \in \mathbb{R}^N$ はそれぞれ結合重みとバイアスのパラメータを表す．また，f は中間層の活性化関数を表し，本研究では Relu 関数 $f(x) = \max(0, x)$ を用いる．一般的に，中間層の次元数 N は入力層の次元数 D よりも小さく設定され，中間層において入力データが次元圧縮される．これらのパラメータは，式 (2) で与えられる入力 $\boldsymbol{x}$ と出力 $\boldsymbol{z}$ の二乗誤差損失を最小にするように確率的降下法 (SGD) を用いて学習する．

$$L(\boldsymbol{x}, \boldsymbol{z}) = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{2} ||\boldsymbol{x}^{(k)} - \boldsymbol{z}^{(k)}||^2\right) \tag{2}$$

学習された自己符号化器の中間層では，次元圧縮された表現 $\boldsymbol{y}^{(k)}$(以下，中間表現) を獲得できる．また，中間層では活性化関数 f によって非線形変換することができ，主成分分析 (PCA) に比べ，複雑な表現を獲得することができる．

2.3　積層ノイズ除去自己符号化器

積層ノイズ除去自己符号化器 (Stacked Denoising Autoencoder; 以下，SDAE)[2] は積層自己符号化器とノイズ除去自己符号化器を組み合わせたもので，より複雑な表現の獲得と汎化性能の向上を図る手法である．

積層自己符号化器 (Stacked Autoencoder) は自己符号化器を階層的に積み重ねた構造を持つ．まず一層目では，入力ベクトル $\boldsymbol{x}^{(k)}$ を用いて自己符号化器を構築する．次に，二層目では，算出された中間層ベクトル $\boldsymbol{y}^{(k)}$ を入力とし，新たな自己符号化器を構築する．以降，自己符号化器の学習を所定の層まで再帰的に行い，積層自己符号化器を構築する．

また，ノイズ除去自己符号化器 (Denoising Autoencoder)[2] は入力に対して確率的にノイズを付与し，そのノイズを除去するように学習を行う．ノイズ付入力ベクトル $\tilde{\boldsymbol{x}}^{(k)}$ の要素は式 (3) で与えられる．

$$\tilde{x}_d^{(k)} = x_d^{(k)} + \epsilon_d \ \ (\epsilon_d \sim \mathcal{N}(0, \sigma^2)) \tag{3}$$

ここで，ϵ_d はノイズであり，平均 0，分散 σ^2 のガウス分布に従うものとする．このノイズ付入力ベクトル $\tilde{\boldsymbol{x}}^{(k)}$ を用いて計算される出力ベクトル $\tilde{\boldsymbol{z}}^{(k)}$ と $\boldsymbol{x}^{(k)}$ の二乗誤差損失が最小になるように式 (4) を用いてパラメータを学習する．

$$L(\boldsymbol{x}, \tilde{\boldsymbol{z}}) = \frac{1}{K} \sum_{k=1}^{K} (\frac{1}{2} ||\boldsymbol{x}^{(k)} - \tilde{\boldsymbol{z}}^{(k)}||^2) \tag{4}$$

3 提案手法

3.1 提案手法の概要

本研究では，一般的なクラスタリング手法がうまく作用しない高次元データを対象とする．このような問題に対し，SDAE による次元圧縮で獲得した中間表現を活用したクラスタリング手法を提案する．

各アイテムの特徴量ベクトルを入力とした SDAE を学習することで，各アイテムに関する特徴量を次元圧縮した中間表現を獲得する．そして，得られた中間表現に対し，k-means 法を適用し，クラスタリングを行う．本手法は，元データに対する k-means 法と比べ，高次元データに対して頑健である．

3.2 提案モデルのアルゴリズム

提案モデルのアルゴリズムを以下に示す．また，図 2 に提案モデルの概要図を示す．

STEP1 SDAE の学習

分析対象データセットを SDAE へ入力し，繰り返し学習を行うことでパラメータを学習し，次元圧縮を行うためのモデルを構築する．

STEP2 中間表現の抽出

構築された SDAE を用い，各アイテムの中間表現を抽出する．

STEP3 中間表現を用いたクラスタリング

STEP2 で抽出された各アイテムの中間表現に k-means 法を適用し，クラスタリングを行う．

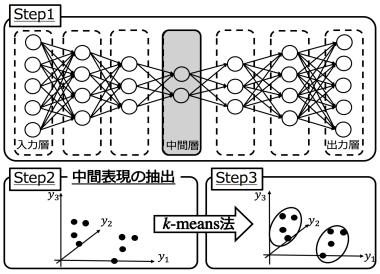

図 2: 提案モデルの概要図

4 分析

4.1 実験条件

提案手法を用いてサイトにおける 2016 年 11 月から 2017 年 9 月の購買履歴データを分析する．データ件数は $270,805(K = 270,805)$ であり，SDAE の入力層と出力層の次元数は表 2 に示す $19,710(D = 19,710)$ を用いた．

本研究では，7 層の SDAE を用いる．中間層の次元数は二層目から順に 5000，1000，200，1000，5000 とした．そして，学習後に得られる第 4 層の 200 次元の中間表現に対し，k-means 法を適用する．また，k-means 法のクラスタ数は $k = 15$ と設定した．

表 2: 分析で用いるデータの変数概要

アイテム情報	説明	ユニーク数
性別	男性，女性	2
小カテゴリ	例)T シャツ，ダウンコート	275
ブランド		20
状態	例)1(最も良い),2(優れている)	5
補助情報	例) 素材，色，サイズ	19,407
割引率		1

4.2 分析結果と考察

クラスタリングを行なった際の各クラスタの所属データ数の結果を表 3 に示す．

表 3: 各クラスタのデータ数

k	データ数	k	データ数	k	データ数
1	11,503	6	10,604	11	51,213
2	14,340	7	15,399	12	14,806
3	15,441	8	20,214	13	35,039
4	24,647	9	11,700	14	1,119
5	16,645	10	16,530	15	11,605

本来比較すべき入力次元数 19,710 の元データに k-means 法を適用した場合，メモリ量の面で実行不可であったため，提案手法は空間計算量の面でも有効と考えられる．

さらに，提案手法の実用性を示すために，特徴的なクラスタを代表的なクラスタとして選び出し，各クラスタに所属するアイテムの特徴を分析する．また，アイテムの全体的な傾向を捉えるため，275 種の小カテゴリを 6 種の大カテゴリにまとめて分析する．ここで，アイテムカテゴリ割合は式 (5) により求める．

$$\text{アイテムカテゴリ割合} = \frac{\text{クラスタ内該当カテゴリのアイテム数}}{\text{クラスタ内総アイテム数}} \tag{5}$$

代表的なクラスタにおけるアイテムカテゴリ割合を表 4 に示す．また，代表的なクラスタにおけるアイテムの状態と割引率に関する分析結果を表 5 にそれぞれ示す．

表 4: 代表的なクラスタのアイテムカテゴリ割合 (%)

k	トップス	ボトムス	ワンピース	アウター	小物類	スーツ
1	**49.21**	25.31	11.06	4.69	9.73	0.01
2	39.32	27.05	13.19	6.11	14.30	0.03
4	34.38	**32.45**	10.15	**14.07**	8.93	0.02
9	39.95	29.46	10.80	6.24	13.54	0.01
14	1.70	1.70	7.15	0.63	**88.83**	0.0

表 5: アイテムの状態と割引率に関する分析結果 (%)

k	状態 1	状態 2	状態 3	状態 4	状態 5	平均割引率
1	1.23	19.60	54.95	**23.53**	0.69	79.34
2	2.73	**34.29**	54.91	7.87	0.20	10.79
4	**12.52**	33.74	36.35	16.96	0.43	0.18
9	3.26	27.93	48.69	**19.60**	0.52	59.71
14	1.88	**40.48**	51.56	5.81	0.27	1.36

表 4 より，アイテムカテゴリ割合に特徴があることが分かる．特に，クラスタ 2 と 9 はそれぞれアイテムカテゴリ傾向が似ている．また，表 5 より，アイテムの状態に特徴が見られる．さらに，クラスタ 1 と 9 は割引率が高く，クラスタ 4 と 14 は低いことが分かる．以上の分析結果をもとに，代表的なクラスタに関して特徴をまとめた結果を表 6 に示す．

表 6: 代表的なクラスタの特徴

k	特徴
1	トップスの割合が比較的高い．状態 4 が比較的多く，高い割引率．
2	状態 2 が比較的多い．低い割引率．
4	アウターやボトムスの割合が比較的高い．状態 1 が特に多く，ほぼ割引なし．
9	状態 4 が比較的多い．高い割引率．
14	小物類の割合が著しく高い．状態 2 は比較的多く，ほぼ割引なし．

クラスタ 2 と 9 は，アイテムカテゴリ傾向は似ていたが，アイテムの状態や割引率に違いが見られた．また，クラスタ 4 は割引率がほぼ 0 であり，中古でもほぼ割引なく購入されるアイテムが所属するクラスタと解釈できる．これは，元々高価だったアイテムが中古として低価格で出品されたことや，商品の状態が良いことから，比較的早期に購入されるためだと考えられる．一方，クラスタ 1 と 14 に関しても，アイテムカテゴリ割合，状態および割引率にそれぞれ特徴があることが分かった．以上のように，クラスタリングによって得られた各クラスタの特徴について，アイテムと割引率の傾向により分析が可能であることが明らかとなった．

本研究では，提案手法によって，高次元データが圧縮され，クラスタリングが可能になったと考えられる．その結果，各クラスタで特徴を分析することができ，マーケティングの施策に応用できると考えられる．

5 まとめと今後の課題

本研究では自己符号化器の中間表現を活用したクラスタリング手法を提案し，サイト A における実際の購買履歴データの分析により，提案手法の有効性を示した．提案手法により，高次元データをクラスタリングすることができ，各クラスタにおけるアイテムの特徴を把握することができた．

また，今後の課題として，販売価格に着目したクラスタリングや本提案手法により得られた知見を用いたマーケティング施策の考案などが挙げられる．

参考文献

[1] G. Hinton and R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, Vol. 313, No. 5786, pp. 504–507, 2006.

[2] V. Pascal, L. Hugo, L. Isabelle, B. Yoshua, M. Pierre-Antoine, “Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion,” *Machine Learning Research*, Vol. 11, pp. 3371–3408, 2010.