5M-09

クラスタリングと確率的主成分分析を用いたアンケートデータ欠損補完に関する検討

福田 智広 † 吉川 大弘 † 古橋 武 †
† 名古屋大学

1 はじめに

企業が顧客の需要や製品に対する評価を把握する方法の 1 つとしてアンケート調査があり，得られたアンケートデータに，多変量解析手法を用いて解析することが多い．しかし，アンケートデータには，回答が未記入の部分が存在する場合があり，そのような欠損を含むデータに対しては多変量解析手法が適用できないため，なんらかの形で欠損部分を補完する必要がある．そこで本稿では，テンソル構造のアンケートデータに対して，回答傾向が類似した回答者をクラスタリングし，クラスタごとに確率的主成分分析と CP(CANDECOMP/PARAFAC) 分解を利用した欠損補完を行う手法を提案する．実際のアンケートデータに提案手法を適用し，その性能評価を行う．

2 提案手法

確率的主成分分析 (Probabilistic Principal Component Analysis : PPCA) は，主成分分析に確率的モデルを適用したものであり，データの潜在変数を用いることで，欠損を補完できる [1]．本稿では，k-means 法を利用して，回答傾向の類似した回答者に分割し，各クラスタに対して CP 分解により 3 階のテンソルを再構築 [2] し，欠損を補完する手法を提案する．なお，以下で単に “テンソル” と表現したときは，すべて 3 階のテンソルを指す．

2.1 CP 分解

代表的なテンソル分解手法に，CP 分解がある．テンソル構造のデータ $\underline{X} \in \mathbb{R}^{n\times m\times l}$ が与えられたとき，CP 分解を用いることで，

$$\underline{X} = \underline{C} \times_1 U_1 \times_2 U_2 \times_3 U_3 \tag{1}$$

と分解できる．ここで，$\underline{C} \in \mathbb{R}^{k\times k\times k}$ は超対角要素が全て 1 のテンソルで，$U_1 \in \mathbb{R}^{n\times k}$，$U_2 \in \mathbb{R}^{m\times k}$，$U_3 \in \mathbb{R}^{l\times k}$ は因子行列と呼ばれる．また，$\times_n$ は n 番目のモードに対するテンソル積で，n-モード積と呼ばれる．CP 分解は

$$\min_{U_1,U_2,U_3} \|\underline{X} - \underline{C} \times_1 U_1 \times_2 U_2 \times_3 U_3\| \tag{2}$$

を計算するものである．

2.2 提案手法

テンソル構造のデータ $\underline{X}$ において，評定尺度法によるアンケートデータでは，n が質問項目数，m が対象項目数，l が回答者数に対応している．そこで本稿では，1 番目のモードを質問モード，2 番目のモードを対象モード，3 番目のモードを回答者モードと呼ぶこととする．

また $\underline{X}$ のモード展開において，質問モード展開では質問項目 ×(対象項目 × 回答者) の行列 $X_{質}$ ができ，対象モード展開では対象項目 ×(回答者 × 質問項目) の行列 $X_{対}$，回答者モード展開では回答者 ×(質問項目 × 対象項目) の行列 $X_{回}$ ができる．提案手法では，k-means 法を用いてクラスタリングした各テンソルに対して，質問および対象モード展開行列に PPCA を適用し，欠損データの補完を行う．その後，補完した各モードを CP 分解の初期値に設定し，データの補完を行う．このとき，クラスタの人数がある閾値より小さい場合は補完を行わず，クラスタ数を 1 減らして，再度クラスタリングを行う．この処理を全回答者の補完が完了するまで行う．

提案手法の手順を以下に述べる．

手順 1: k-means 法を利用して，テンソル (アンケートデータ) をクラスタリングする．このとき，欠損部分は各質問項目の平均値で埋める．各クラスタに以下の操作を行う．

手順 2: クラスタの人数が閾値以上の場合は手順 3 以下の通り補完を行う．全クラスタを確認し，補完ができていない回答者がいれば，手順 1 へ戻る．このときクラスタ数を 1 つ減らす．

手順 3: モード展開した行列 X_n において，欠損部分がない行列 X_{nObs}(観測部分と呼ぶ) と欠損部分

A Study on Imputation of Missing Value in Questionnaire Data Using Clustering and Probabilistic Principal Component Analysis
Tomohiro Fukuta , Tomohiro Yoshikawa and Takeshi Furuhashi
†Nagoya University

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.

がある行列 X_{nMiss} に分ける．

$$X_n = [X_{nObs}, X_{nMiss}] \tag{3}$$

手順 4: 観測部分 X_{nObs} に高階特異値分解 ($Higher\ Order\ Singular\ Value\ Decomposition$: $HOSVD$) を適用し，U_{nObs} を求める．

手順 5: 手順 4 で求めた U_{nObs} を用いて，各モードにおける潜在変数 z を求める．

$$z = (U_{nObs}^T U_{nObs})^{-1} U_{nObs}^T X_{nMiss} \tag{4}$$

手順 6: 手順 5 で求めた z を用いて，補完値を求める．

$$X_{nMiss} = U_{nObs} z \tag{5}$$

手順 7: 手順 5，6 で補完した各モード展開行列を評点 (1〜5) に規格化したのち，補完した各モードに $HOSVD$ を適用し，U_n を求める．

手順 8: 手順 7 で求めた U_n を CP 分解の初期値にし，テンソル再構築を行うことで，欠損部分が補完されたテンソル $\underline{X_{imp}}$ が求まる．

$$\underline{X_{imp}} = \underline{C} \times_1 U_1 \times_2 U_2 \times_3 U_3 \tag{6}$$

手順 9: 欠損を含む全ての回答者を補完していないならば，手順 2 へ戻る．全ての補完値が求まったら，1 つのテンソルにまとめ，補完を完了する．

3 実験

3.1 実験方法

5 段階評定尺度法による実際のアンケートデータに対して，欠損部分以外の平均値を用いて補完する従来手法 [3]，CF(協調フィルタリング) 法，KNN(K 近傍) 法，CP-CF 法 [4]，および提案手法を適用し，欠損補完に対する精度の比較を行った．このとき，提案手法のクラスタ数を 10，ユーザ閾値を 200 とした．ここで，欠損を行う回答者を，ランダムに抽出された回答者全体の 10%〜50% の 10% 刻みとし，抽出された各回答者の欠損割合は 10% とした．評価指標には，評点の傾向を捉えることができたかを示す評点傾向を用いた．

3.2 結果と考察

各手法による補完精度を図 1 に示す．提案手法はどの欠損率においても，他の手法よりも精度よく補完できていることがわかる．しかし，欠損率が高くなるにつれて，CP-CF 法の補完精度との差が減っている．これは，欠損率が高くなるにつれて，使用できる観測部分の情報が減り，クラスタリング精度と補完精度がともに悪くなってしまったためであると考えられる．

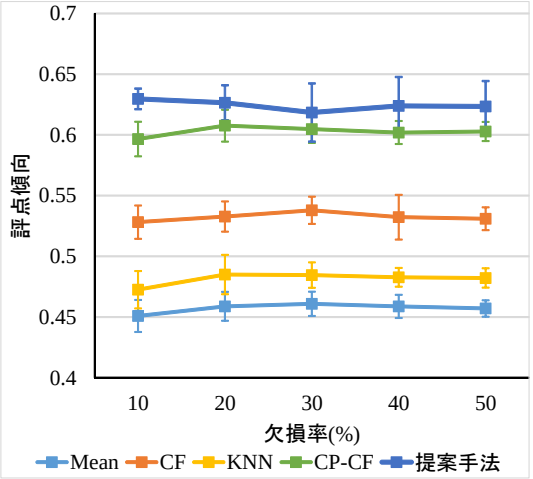

図 1: 補完精度

4 おわりに

本稿では，3 階のテンソル構造のアンケートデータに対し，クラスタリングを適用した上で，確率的主成分分析と CP 分解を用いて欠損補完を行う手法を提案した．実際のアンケートに適用し，欠損率に関わらず，提案手法が従来手法よりも補完精度が高いことを示した．今後の課題として，欠損を含むデータのクラスタリングに対する検討，ランダムではなく意図的に作成された欠損補完に対する検討などが挙げられる．

参考文献

[1] L.Qu et al. : PPCA-based missing data imputation for traffic flow volume: a systematical approach, IEEE Trans, Intelligent Transportation Systems, vol.10, pp.512-522, 2009.

[2] T.G.Kolda : Multilinear operators for higher-order decompositions. Technical Report SAND2006-2081, Sandia National Laboratories, Albuquerque, NM and Livermore 2006.

[3] I.Myrtveit, E.Stensrud, U.H.Olsson : Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods, IEEE Trans, Software Engineering, vol.27, pp.999-1013, 2001.

[4] L.Garg et al. : Tensor-Based Methods for Handling Missing Data in Quality-of-Life Questionnaires, IEEE Journal of Biomedical and Health Informatics, vol.18, pp.1571-1580, 2014.

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.