時系列データの逐次補完による類似度行列作成手法の提案

1X13C092-9　服部 達也
指導教員　後藤 正幸

1　研究背景と目的

近年の IoT 技術の発展に伴い，企業にはインターネットを介して自社製品を利用する顧客の日々の利用履歴データを収集する環境が整いつつある．そのデータ活用の 1 つに，顧客の製品の利用傾向の分析がある．利用傾向には，利用頻度が減少していく離反傾向や，安定利用や増加傾向などの継続的な利用が続く優良な利用傾向などがある．顧客毎の利用傾向を分析し，得られた結果を収益の向上を目的とした様々な施策へと結び付けることが期待できる．特に，新規に獲得した顧客（以下，新規顧客）の利用傾向の分析及び施策の策定は今後の収益向上のために重要な課題となる．

本研究ではこのような観点に基づき，新規顧客の特徴を事前に把握するため，利用傾向が類似している既存顧客を推定することを考える．これにより，新規顧客に対して類似度の高い既存顧客と同様の施策が適用可能となる．しかし，新規顧客に対しては短期間の利用履歴データしか得られていないため，この点を考慮に入れた顧客間類似度の推定法を構成する必要がある．

そこで本研究では，既存顧客のデータ（完全データ）に対し，新規顧客のデータを不完全データとみなし，完全データと不完全データをまとめてデータ間の類似度を算出する新たな枠組みを構築する．通常，完全データ同士の類似度は推定精度が高く，欠損が増えるほど精度が下がると考えられる．このため，同じ欠損であっても，比較的精度が良く補完ができそうな箇所とそれが難しい箇所とが混在している．これらを一律に欠損値補完しようとすると，全体として精度が下がってしまう可能性がある．そこで本研究では，製品の利用履歴データの欠損パターンに着目し，不完全データのうち欠損の少ないデータから順に，類似度に基づいたデータの補完と，類似度の再計算を交互に繰り返す方法を提案する．この方法により，利用可能なデータを最大限活用し，逐次的な補完によって信頼性の高い欠損値補完を行うことで，類似度行列全体の精度向上が期待できる．提案手法の有効性を示すために人工データに提案手法を用いた実験を行う．最後にグローバルにプリンタ・複合機事業を展開している印刷機器メーカー提供のプリンタ利用履歴データに適用し，分析を行う．

2　準備

2.1　問題設定

本研究で対象としているプリンタの利用履歴データは，印刷枚数などの履歴が月次で蓄積された時系列データである．過去一定期間でデータが観測されている顧客を既存顧客，データの観測が途中から開始されている顧客を新規顧客とする．一般的に製品の利用傾向は利用開始からの経過月数に依存する．そこで，顧客ごとに利用開始からの月数で時系列データを捉えることにより，データ全体の形状は図 1 のように与えられる．既存顧客のデータを完全データ，新規顧客のデータを不完全データとみなし，観測されていないデータ，すなわち新規顧客がまだ製品を購入していない期間のデータを欠損として扱う．また，製品の利用傾向には図 2 に示したように安定，増加，減少，離反などが考えられる．

2.2　類似度行列と欠損処理

類似度行列は，2 つのサンプル（顧客データ）間の類似度を要素に持つ対称行列である．サンプル間の類似度を測る指標にはユークリッド距離やマハラノビス距離など，様々に存在するが，本研究では，データの規模を考慮せず，サンプル間の利用傾向の類似性を測りたいという観点から，コサイン類似度を用いる．

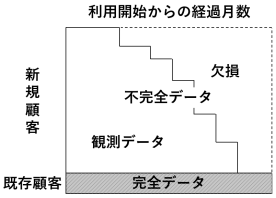

図 1. データの形状

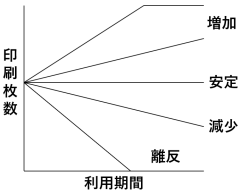

図 2. 利用傾向の例

また，データに欠損を含むサンプル間の類似度を算出する際，予め欠損に対して処理を行う必要がある．欠損の処理に関する既存研究は大きく分けて 2 つある．1 つはリストワイズ法と呼ばれる方法である．これは，欠損値を含むサンプル，または変数を削除し，データに欠損を含まない完全なサンプルのみで分析を行う方法である．もう 1 つは欠損値を推定し，補完する方法である．代表的な推定方法として，平均値推定法が挙げられる．

3　提案手法

3.1　概要

欠損データに対する従来手法であるリストワイズ法や平均値推定法は，欠損がデータ中にランダムに生じるようなデータに対して有効である．しかし本研究では，図 1 に示したように欠損は観測期間中でランダムに生じるものではなく，連続して発生するようなデータを対象としている．

そこで本研究では，対象データの欠損パターンを考慮した欠損値補完法を考える．特に，対象とするデータは系列の途中で欠損が生じないという性質を活かし，利用可能なデータを最大限活用し，より信頼性が高い欠損値補完を行うことを考える．長い系列の不完全データから順次欠損の補完を行うことで，短い系列の新規顧客データに対しても，信頼性を維持した下での欠損値補完および類似度算出を行う手法を提案する．具体的には，まず完全データを用いて最も系列の長い不完全データの欠損の補完と類似度の算出を行う．その後，完全データと欠損が補完された不完全データを併せて新たな完全データとみなし，次に系列の長い不完全データの欠損の補完と類似度の算出を行う．この操作をすべての不完全データに対して順に実行し，最終的な類似度行列を求める．これにより，完全データの情報を不完全データへ反映させた欠損値補完が実現し，高い精度の類似度算出が期待される．

3.2　提案アルゴリズム

N 個のサンプルのうち，n 番目のサンプル系列を $\boldsymbol{x}_n$，tヶ月目の要素を $x_{nt}(t = 1, 2, ..., T)$ とする．ただし，データが欠損しているとき $x_{nt} =$ ‘$null$’ とする．$\boldsymbol{x}_n$ の要素のうち ‘$null$’ ではない要素の数を $l(\boldsymbol{x}_n)$ としたとき，$l(\boldsymbol{x}_n) = T$ となるサンプルが既存顧客，$l(\boldsymbol{x}_n) < T$ となるサンプルが新規顧客である．また，$l(\boldsymbol{x}_n)$ の最小値を l_{min} とし，$\boldsymbol{x}_n, \boldsymbol{x}_m$ 間の $1 \sim t$ 期までの特徴量によるコサイン類似度を $cos_t(\boldsymbol{x}_n, \boldsymbol{x}_m)$ と表す．$\mathcal{D}_i$ は $l(\boldsymbol{x}_n) = i$ となる $\boldsymbol{x}_n$ の集合を示す．また，$N \times N$ 類似度行列 $\boldsymbol{S}$ の n 行 m 列の成分を s_{nm} とする．

提案アルゴリズムは以下の 3 つの考え方に従っている．(1) データが多い（欠損の少ない）サンプルほど情報が多く，算出される類似度が信頼できるため，欠損の少ないサンプルから順に，完全データのサンプルとの類似度算出を行う．(2) 欠損が後半に集中する時系列データであるため，補完の際，

全ての欠損を一度に補完せず，観測データに近い箇所の欠損から順に補完することで，信頼度の高い順に補完を行う．(3) より信頼できる箇所から欠損値補完を繰り返すことで，信頼性の低い補完値の影響が他の欠損値補完に伝播することを防ぎ，推定された類似度の精度を向上することができる．以下に厳密な提案アルゴリズムを擬似コードで示す．

Algorithm 提案アルゴリズム

for $i = l_{min}$ to T **do**
  $\mathcal{D}_i = \phi$
for $n = 1$ to N **do**
  $\mathcal{D}_{l(\boldsymbol{x}_n)} \leftarrow \boldsymbol{x}_n$
for $i = l_{min}$ to T **do**　＃欠損の数が同じデータ間の
  for each $\boldsymbol{x}_n \in \mathcal{D}_i$ **do**　類似度はそのまま計算する．
    for each $\boldsymbol{x}_m \in \mathcal{D}_i$ **do**
      $s_{nm} = cos_i(\boldsymbol{x}_n, \boldsymbol{x}_m)$
for $i = T-1$ to l_{min} **do**　＃欠損の少ない順に行う．
  for $j = i+1$ to T **do**
    for each $\boldsymbol{x}_n \in \mathcal{D}_i$ **do**
      $sum = 0, x_{nj} = 0$
      for each $\boldsymbol{x}_m \in \mathcal{D}_T$ **do**
        $s_{nm} = cos_{j-1}(\boldsymbol{x}_n, \boldsymbol{x}_m)$　＃類似度算出
        $sum = sum + s_{nm}$
        $x_{nj} = x_{nj} + x_{mj} \times s_{nm}$
      $x_{nj} = x_{nj}/sum$　＃欠損値補完
  for each $\boldsymbol{x}_n \in \mathcal{D}_i$ **do**　＃全ての欠損を補完した後，
    for each $\boldsymbol{x}_m \in \mathcal{D}_T$ **do**　再度類似度を更新
      $s_{nm} = cos_T(x_n, x_m)$
  $\mathcal{D}_T \leftarrow \mathcal{D}_T \cup \mathcal{D}_i$　＃補完済みの新規顧客を既存顧客とする．

4　人工データを用いた実験

4.1　実験条件

提案手法の有効性を検証するため，人工データを用いた実験を行う．人工データは研究対象データに基づき月次データとして考え，図 2 に示したような特性（1．増加した後安定，2．単調増加，3．一定，4．単調減少，5．減少した後 0 となる）を持つ 5 つのクラスタからなるデータを想定する．

初月の印刷枚数を 4,800 とし，各特性に基づき関数を定義する．特性 1 は 12ヶ月目まで単調増加した後一定となる関数，特性 2，3，4 はそれぞれすべての期間で単調増加，一定，単調減少である関数，特性 5 は 12ヶ月目で 0 となり以降 0 となる関数で表す．以上の関数で与えられる値に，$N(0, 800^2)$ に従う正規ノイズを加え，各月の利用履歴人工データを作成する．各クラスタにつき 2,100 件ずつ，合計 10,500 件，17ヶ月分のデータを作成した．

一般的に製品の利用は 3ヶ月で安定すると言われることから，3ヶ月以上データが存在するサンプルのみを対象とする．また，17ヶ月存在しない新規顧客データについては，5 %刻みで 5〜40 %のデータを削除し，欠損率毎にデータセットを作成する．欠損させる前のデータで算出した類似度行列を真とし，推定誤差を平均二乗誤差（RMSE）で評価する．

比較手法については，2 つのサンプル間の共通している月のデータで類似度を算出するリストワイズ法を比較手法 1 とする．欠損を各サンプルの観測されたデータの平均値で補完し類似度を算出する平均値推定法を比較手法 2 とする．

4.2　実験結果と考察

いずれの欠損率においても提案手法が比較手法よりも高い精度（RMSE）を示し，提案手法の有用性が確認できた．

比較手法 1 と提案手法では，いずれの欠損率でも提案手法の精度が高かった．これは，比較手法 1 では欠損を持つサンプルの類似度算出の際に一部のデータを無視するため，適切な類似度が得られなかったと考えられる．一方で，比較手法 2 と提案手法では，欠損率が低いときはほぼ同等であるが，欠損率が高くなるにつれて提案手法が優れる結果となった．これは，データ特性として月の経過による増加や減少を持たせているため，平均値推定法では，欠損率の増加に伴って補完された平均値との誤差が大きくなり，類似度の推定誤差が大きくなったためであると考えられる．一方で提案手法では，精度の高い補完ができる箇所から優先的に埋めていくことで，信頼性の高い値を次の欠損値補完に用いる方法となっている．これにより，信頼性の低い補完値が他に影響することを防いでいる．このため，欠損率の増加の影響を比較的受けにくい手法であると考えられる．

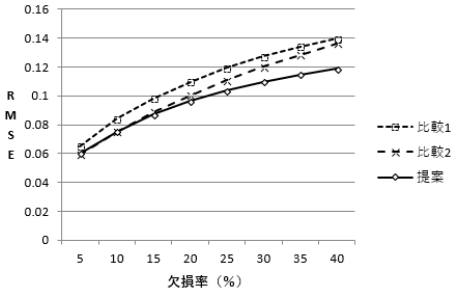

図 3. 欠損率による RMSE の変化

5　実データを用いた実験

5.1　実験条件

本研究で用いる実データは 2016 年 1 月から 2017 年 5 月までの 17ヶ月間に蓄積された，プリンタの月毎の印刷枚数である．利用期間が 6ヶ月までの顧客を新規顧客と定義する．全顧客数は 4,055，そのうち新規顧客数は 1,349 である．

本実験では，実データに対して提案手法を適用し，得られた類似度行列を用いてクラスタリングを行う．クラスタリングはウォード法を用いた階層クラスタリングを行い，クラスタ数を 5 とした．

5.2　実験結果と考察

表 1. 各クラスタの新規顧客の割合

	顧客数	新規顧客数	新規顧客数の割合
クラスタ1	1659	535	0.3225
クラスタ2	698	350	0.5014
クラスタ3	677	71	0.1049
クラスタ4	419	55	0.1313
クラスタ5	602	338	0.5615
合計	4055	1349	0.3327

提案手法により算出した類似度行列を用いたクラスタリングを行うことで，各クラスタ内の既存顧客の特徴により新規顧客の分析を行うことが可能となった．提案手法が想定したデータに対して有効であることが分かる．

6　まとめと今後の課題

本研究では，既存顧客を完全データ，新規顧客を不完全データとみなし，データの形状に着目した，不完全データの逐次補完による顧客間の適切な類似度の算出方法を提案した．人工データを用いた実験により，提案手法の有用性が確認された．また，実データを用いた実験では実際のデータに提案手法を適用し，算出した類似度をもとにクラスタリングを行い，新規顧客と既存顧客を合わせたデータ間の類似度を適切に算出することが確認できた．今後の課題として，次元数やデータ数を増やした際の計算時間の短縮などが挙げられる．

参考文献

[1] 真田祐希，大井貴裕，石田崇，後藤正幸，“欠損値を含むデータのクラスタリングのための Random Forest を用いた類似度算出法，”電子情報通信学会論文誌 (D)，vol.J97-D，No.1，pp.239-243, 2013.

[2] MARK HUISMAN，“Imputation of Missing Item Responses:Some Simple Techniques，”Quality & Quantity 34，pp.331-351，2000.