欠損を含むデータのクラスタリングに適した Random Forest による類似度算出法

1X10C036-4　木村美月
指導教員　後藤正幸

1　研究背景と目的

近年，情報技術の発達に伴い，データ分析に基づく経営・マネジメントの重要性が増しており，そのためのデータ分析手法の一つとして，クラスタリングが挙げられる．クラスタリングでは，サンプル間の類似度を測り，類似度の大きいサンプル同士をクラスタに併合することでグルーピングを行う．しかし，実データにはしばしば欠損があり，そのような欠損を含むデータからは直接類似度を算出することができないため，何らかの工夫が必要となる．

欠損を含むデータに対するクラスタリングのための類似度算出法の一つに，真田らの手法 [1] がある．真田らの手法では，欠損を含まない部分集合（以下，縮退行列）を複数組ランダム抽出し，観測値のみからなる各縮退行列から Random Forest[2] を用いて求めた類似度をアンサンブルすることで全サンプル間の類似度を求める．その際，あるサンプル間について観測データの共起がみられない場合は，それらの類似度を求めることが出来ない．そこで，真田らの手法では欠損値補完法を用いた例外処理によって，縮退行列からは算出不可能な類似度を補っている．しかし，欠損率が高くなるにつれ算出不可能な類似度が増加するため，クラスタリング精度が欠損補完値のバイアスの影響を受けてしまう．

そこで，本研究では，真田らの手法において観測値のみから求めた類似度の推定精度は高い点に着目し，推定された類似度で重み付けをした平均値の補完と類似度算出を繰り返す新たな手法を提案する．具体的には，観測値のみから求めた類似度を用いて欠損データの部分補完を行い，サンプル間の観測値の共起を増やすことで，類似度算出が可能なサンプルの組を増加させる．１回の施行では全サンプル間の類似度が算出できないため，類似度算出と類似度による補完を繰り返し，収束させるアルゴリズムとなる．提案手法の有効性を示すために，ベンチマークデータを用いた実験を行う．

2　準備

2.1　問題設定

欠損を含む $N \times M$ のデータ行列 $\boldsymbol{D} = [x_{i,j}]$ が与えられているものとする $(1 \leq i \leq N, 1 \leq j \leq M)$．ここで，$x_{i,j}$ は i 番目のサンプル（サンプル i と呼ぶ）の j 番目の変数の値であり，$x_{i,j} \in \mathcal{R} \cup \{\phi\}$ である．ただし，$\mathcal{R}$ は実数全体の集合，ϕ はデータの欠損を示す．

2.2　Random Forest

Random Forest [2]（以下 RF）は複数の決定木を用いたアンサンブル学習器である．RF では，学習データから T 個のブートストラップサンプルを生成し，M 次元の各ブートストラップサンプルに対してランダムに選択した m 個 $(m < M)$ の変数を用いて決定木を構築する．このとき，サンプル間の類似度は，各決定木においてサンプル同士が同じ葉ノードに属する回数を決定木の数 T で割る事で算出される．この類似度は一般的な距離尺度であるユークリッド距離とは全く異なる性質を持ち，精度の高いクラスタリングが可能になる [1]．

2.3　真田らの類似度算出法

真田らの手法は，欠損を含むデータから類似度を算出するためのアンサンブル手法であり，大きく三つのステップで類似度行列を求める．はじめに，入力データ行列 $\boldsymbol{D}$ から，欠損値を含まない縮退行列を複数ランダム抽出する．次に，各縮退行列に RF を適用して類似度行列を算出する．最後に，得られた複数の類似度行列を統合して一つの類似度行列を求める．

$k = 1, 2, \cdots, K$ に対し，縮退行列 $\boldsymbol{D}_k$ は，変数選択と欠損を含むサンプルの削除によって生成される．はじめに，データ行列 $\boldsymbol{D}$ から Q 個 $(Q < M)$ の変数をランダムに選択することで $N \times Q$ の部分データ行列を抽出する．次に，部分データ行列から欠損を含むサンプルを除いて $N_k \times Q$ の行列を生成する $(N_k \leq N)$．これを K 回繰り返すことで，K 個の縮退行列 $\boldsymbol{D}_k$ を求める．縮退行列 $\boldsymbol{D}_k$ に RF を適用してサンプル間の類似度を算出し，$\boldsymbol{D}_k$ に含まれるサンプル間での類似度行列 $\boldsymbol{S}_k$ を求める．統合類似度行列 $\boldsymbol{U} = [u_{i,i'}] \in \mathcal{R}^{N \times N}$ は，K 個の類似度行列 $\boldsymbol{S}_k$ を平均統合することで求められる．$\boldsymbol{U}$ の要素 $u_{i,i'}$ は，各 $\boldsymbol{S}_k$ におけるサンプル i とサンプル i' の類似度の総和を，K 個の $\boldsymbol{S}_k$ のうちサンプル i とサンプル i' が共起した回数で割ることで求める．

ここで，統合類似度行列 $\boldsymbol{U}$ の要素 $u_{i,i'}$ のうち，K 個の類似度行列 $\boldsymbol{S}_k$ において一度も算出されていないサンプル間の類似度は求めることができない．全サンプル間の類似度が求められていない場合，クラスタリングを行うことができないため，例外処理を行う必要がある．例外処理では，データ行列 $\boldsymbol{D}$ に対し平均値推定 [3] による欠損値補完を行い，補完後のデータ行列に RF を適用して求められたサンプル間の類似度を統合類似度行列 $\boldsymbol{U}$ に代入している．

3　提案手法

3.1　概要

真田らの手法では，縮退行列から求められなかったサンプル間の類似度を最後に平均値推定によって補完したデータから求めているため，欠損率が高くなるにつれてクラスタリング精度が補完値のバイアスの影響を強く受けてしまう．したがって，観測データの情報を用いて補完値のバイアスを修正する必要がある．そこで，本研究では，RF を用いた類似度算出と類似度を用いた重み付け平均値による欠損値補完の繰り返し処理を行う類似度算出法を提案する．これにより，欠損値補完に観測データから求めた類似度の情報が反映され，平均値補完による例外処理よりも精度の高い類似度の推定が期待でき，クラスタリング精度の向上が見込まれる．提案手法の概要と真田らの手法との関係を図 1 に示す．

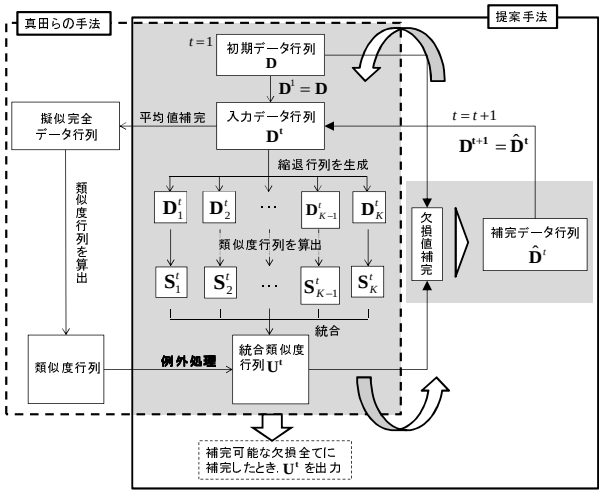

図 1．提案手法の概要

3.2　類似度算出

類似度算出では，真田らの手法と同様に，アンサンブルにより統合類似度行列を求める．提案手法では，欠損を含

むデータ行列を初期データ行列 $\boldsymbol{D}^1 = \boldsymbol{D}$ とし，次のように類似度算出と欠損値補完を繰り返し行う．t ステップ目 $(t = 1, 2, 3, \cdots)$ の入力データ行列を $\boldsymbol{D}^t = [x^t_{i,j}]$ として，K 個の縮退行列 $\boldsymbol{D}^t_k = [(\tilde{x}^t_k)_{i,j}]$ をランダムに生成し，類似度行列 $\boldsymbol{S}^t_k = [(s^t_k)_{i,j}]$ を推定して $(1 \leq k \leq K)$，統合類似度行列 $\boldsymbol{U}^t = [u^t_{i,i'}]$ を生成する．$\boldsymbol{U}^t$ を用いて欠損値を補完したデータ行列 $\hat{\boldsymbol{D}}^t = [\hat{x}^t_{i,j}]$ を次の入力データ行列 $\boldsymbol{D}^{t+1} = \hat{\boldsymbol{D}}^t$ とする．

ただし，統合類似度 $\boldsymbol{U}^t$ において，K 個の類似度行列 $\boldsymbol{S}^t_k$ で一度も算出されていないサンプル間の類似度は 0 とする．

3.3 欠損値補完

欠損値補完では，各ステップ t において，初期データ行列 $\boldsymbol{D}$ の全ての欠損に対して，統合類似度行列 $\boldsymbol{U}^t$ を用いて補完を行う．$x_{i,j}$ が欠損しているとき，変数 j についてデータが観測されている全てのサンプル i' の観測値 $x_{i',j}$ をサンプル i との類似度 $u^t_{i,i'}$ によって重み付けした平均値で補完する．類似度 $u^t_{i,i'}$ は観測値を用いて算出されているため，精度の高い欠損値補完が期待できる．ステップを重ねることで類似度行列が更新されていくため，欠損への補完値も最新値に更新されるものとする．

t ステップ目における補完値を式 (1) に示す．

$$\hat{x}^t_{i,j} = \begin{cases} \frac{\sum^N_{i'}(I(x_{i',j}) \times x_{i',j} \times u^t_{i,i'})}{\sum^N_{i'}(u^t_{i,i'} \times I(x_{i',j}))} & x_{i,j} = \phi \\ x_{i,j} & x_{i,j} \neq \phi \end{cases} \quad (1)$$

式 (1) における $I(x_{i,j})$ はインジケータ関数であり，$x_{i,j}$ の値が欠損しているとき 0 をとり，それ以外のときは 1 をとる．

補完済みデータ行列 $\hat{\boldsymbol{D}}^t$ を次の類似度算出における入力データ行列 $\boldsymbol{D}^{t+1}$ とする．これを，補完数が収束するまで繰り返す．

3.4 提案アルゴリズム

以下に提案手法のアルゴリズムを示す．

(1) $t = 1$ とし，入力データ行列 $\boldsymbol{D}^1 = \boldsymbol{D}$ と設定する．

(2) 入力データ行列 $\boldsymbol{D}^t$ から K 個の縮退行列 $\boldsymbol{D}^t_k$ を生成する．各縮退行列 $\boldsymbol{D}^t_k$ に対して RF を適用し，K 個の類似度行列 $\boldsymbol{S}^t_k$ を得る．

(3) K 個の類似度行列 $\boldsymbol{S}^t_k$ を，一つの統合類似度行列 $\boldsymbol{U}^t$ に統合する．そのとき，K 個の類似度行列 $\boldsymbol{S}_k$ の中で一度も共起のなかったサンプル i と i' の類似度は 0 とする．

(4) 初期データ行列 $\boldsymbol{D}$ の欠損に対して (1) 式を用いて補完し，$\hat{\boldsymbol{D}}^t$ を算出する．

(5) $\boldsymbol{D}^{t+1} = \hat{\boldsymbol{D}}^t$，$t = t+1$ として (2)〜(4) を繰り返し，補完数に変化がなくなったら統合類似度行列 $\boldsymbol{U}^t$ を出力して終了． □

4 実験と結果

4.1 実験条件

実験では，UCI 機械学習レポジトリよりデータセット WINE を用いた．データのサンプル数は $N = 178$，変数の数は $M = 13$，カテゴリ数は $C = 3$ であり，本実験でのクラスタリングにおけるクラスタ数も $L = 3$ とした．選択する変数は $Q = 3$ とし，縮退行列数は $K = 100$ とした．RF における決定木の生成法は CART [2] を用い，木の数は 500，分岐の際に選択する変数は 1 とした．また，クラスタリング手法は階層クラスタリング手法であるウォード法を用いた．完全データを欠損データとするために元のデータを 10〜70 %欠損させ，各欠損率について欠損場所をランダムに変えて 100 回繰り返し，その平均を結果とした．比較手法として真田らの手法で例外処理の補完法に平均値推定を用いる方法（mean）と MICE[3] を用いる方法（MICE）の 2 手法を用いた．MICE は，一般に補完値の推定精度が非常に高いといわれている欠損値補完法である．

4.2 評価方法

クラスタリングの性能の評価方法として一般的に用いられる指標であるエントロピーを用いた．エントロピー E は，$\mathcal{C} = \{C_1, C_2, ..., C_L\}$ をクラスタリング結果，$\mathcal{A} = \{A_1, A_2, ..., A_L\}$ を正解となるクラスタリング結果，$E(A_i)$ を正解クラスタリング結果が A_i のときのエントロピーとするとき，式 (3) で定義される．

$$E = \sum_{i=1}^{L} \frac{\sum_{j=1}^{L} | C_i \cap A_j | \times E(A_i)}{N} \quad (2)$$

ただし，$|\cdot|$ は集合の要素数を表す．E の値は 0 から 1 の間をとり，値が小さいほどクラスタリング結果が良好であることを意味する．

4.3 結果と考察

実験結果を図 2 に示す．

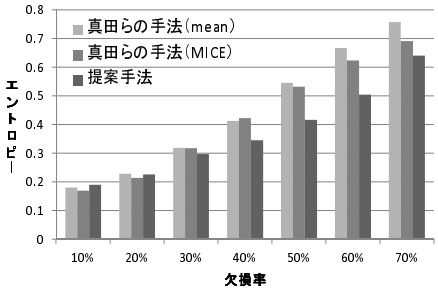

図 2. クラスタリング結果

図 2 より，欠損率 30 %以上では提案手法のクラスタリング精度が真田らの手法よりも優れている．欠損率 20 %以下では，真田らの手法と提案手法で差は微小であるが，これは算出不可能な類似度の数が少なく，クラスタリング精度に対する欠損値補完の影響が小さいためであると考えられる．提案手法では，観測データから推定された精度の高い類似度を用いて欠損値補完し，次のステップでの類似度算出に活用することにより，より多くの観測値を類似度算出に活用できたため，クラスタリング精度が向上したと考えられる．さらに，真田らの手法（MICE）よりも提案手法のクラスタリング精度が優れていることから，例外処理における欠損値補完の精度を上げようとするよりも，途中までに得られている類似度を活用する提案手法の方が精度の高い類似度を得られるといえる．

5 まとめと今後の課題

本研究では，欠損を含むデータに対するクラスタリングに適した類似度算出法として，類似度による欠損値補完と補完したデータによる類似度算出の繰り返しアルゴリズムによる方法を提案した．数値実験により，提案アルゴリズムの有効性が確認された．今後の課題として，欠損率を考慮した適切な縮退行列次元数 Q の決定法などが挙げられる．

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] L.Breiman，“Random Forests,” *Machine Learning, 45*, pp.5-32, 2001.

[3] A.C.Acock，“Working with Missing Values,” *Journal of Marriage and Family*, vol.67, no.4, pp.1012-1028, Nov. 2005.