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

真田　祐希[†]　　大井　貴裕[†]
石田　　崇[††]（正員）　　後藤　正幸[†]（正員）

Similarity Matrix Based on Random Forest for Clustering of Incomplete Data

Yuki SANADA[†], Takahiro OI[†], *Nonmembers*,
Takashi ISHIDA[††], and Masayuki GOTO[†], *Members*

[†] 早稲田大学創造理工学部経営システム工学科，東京都
Faculty of Creative Science and Engineering, Waseda University, Tokyo, 169–8555 Japan
[††] 早稲田大学メディアネットワークセンター，東京都
Media Network Center, Waseda University, Tokyo, 169–8050 Japan

あらまし　本論文では，欠損値を含む多変量データから，クラスタリングのための類似度行列を生成する方法について検討する．ランダムに縮退させた部分完全データ行列を生成し，ランダムフォレストを用いて類似度を算出する操作を繰り返して混合する方法を示す．

キーワード　ランダムフォレスト，欠損データ，クラスタリング，類似度行列

1. まえがき

近年，優れた予測性能をもつアンサンブル手法の一つとして Random Forest（RF）[1], [2] が注目されている．RF は高速かつ高精度でデータの分類・回帰が可能なアンサンブル予測器であり，精度のよい類似度行列算出法としても応用可能である [3]．RF は目的変数やクラスラベルのないデータからの “教師なし学習” における類似度算出法についても適用可能であり，R による randomForest パッケージでも実装されている [4]．RF を用いた学習データ間の類似度算出法の有用性は幾つかの研究で指摘されており [5]，算出された類似度をクラスタリングに用いることができる．

一方，クラスタリングでは，欠損のない完全データのみが用意されればサンプル間の距離の算出は容易であるが，実データではしばしば欠損が含まれることもある [6]．データに欠損が含まれる場合には直接，距離を計算することが不可能となるため，欠損値を含むデータから類似度を推定する方法が必要である．欠損処理に対する最も簡便な方法は，欠損値を平均値などで補完（Imputation）して類似度を計算する方法である [7], [8]．しかし，単純な補完法では欠損率が高くなるにつれてより多くの補完データが用いられるため，これらの補完データから算出された類似度は，完全データから算出された類似度と比べ，相対的に精度が劣化してしまう．そのため，完全情報最尤推定法（Full Information Maximum Likelihood: FIML）や多重代入法（Multiple Imputation）といった方法が提案されている [7], [9]．しかし，目的をクラスタリングのための類似度算出に限った場合，確率モデルを陽に定義しなければゆう度が計算できない完全情報最尤推定法は適さない．多重代入法も，欠損値をランダムに補完するような方法であれば適用可能であるが，MCMC 法（Markov Chain Monte Carlo Method）[9] や Chained Equations [10] を用いるような系統的な方法は，回帰モデルなどの確率モデルを仮定した上で事後予測分布を計算して乱数生成する必要がある．クラスタリングを目的としたサンプル間の類似度算出では，陽に確率モデルを限定する必要のない，汎用的な推定手法であることが望ましい．しかし，RF を用いた類似度算出法における欠損値の取り扱いはアドホックな代入法によるものが多く [11]，改善の余地があると考えられる．

そこで本研究では，観測されたデータのみから推定された類似度行列の精度は高いことに着目し，欠損を含まない箇所を抽出した部分集合データから得られる類似度行列を活用する方法を考える．しかし，そのような部分集合データの抽出の仕方は一意ではなく，全てのサンプル間の類似度が得られる保証もないため，何らかの工夫が必要である．そこで本論文では，ランダムに抽出した欠損のない部分集合データから類似度を算出する操作を繰り返し，得られた結果を統合して類似度を算出するアンサンブル手法を提案する．提案手法の有効性を示すため，ベンチマークデータを用いた実験を行い，類似度の推定精度とクラスタリングの精度の両面から評価を行う．

2. 準　　備

2.1 Random Forest

Random Forest（以下，RF）[12] は，学習データから生成した T 個のブートストラップサンプルに対して，それぞれ独立に T 個の決定木を構築する．決定木の構築の際に，全 M 個の変数の中からランダムに選択された m 個（$m < M$）の属性変数を用いる点が特徴である．RF により予測を行う際には，対象のデータを学習で構築された T 個の決定木に入力し，各決定木から得られた結果を統合することで最終的な出力結果を得る．RF は，このように生成された複数の決定

木を用いることで，過学習を防ぎ，高い汎化性が得られることが知られている．

2.2 RF を用いた教師なし類似度行列算出法

類似度行列は，二つのサンプル間の類似度を要素にもつ対称行列である．各要素は 0 から 1 の間の値を取り，1 に近いほどデータ同士が類似していることを表す．また，行列の対角要素は全て 1 となる．

目的変数やクラスラベルをもたないデータに対して，RF を用いた教師なし学習を行うことによって，サンプル間の類似性を表す類似度行列を得ることができる [4], [5]．具体的には，得られている教師なしデータ（学習データ）を “クラス 1” とし，これに対して “クラス 2” に属する人工データを，各変数が学習データの経験周辺分布に従うように乱数で発生させる [5]．RF によって，これら二つのクラスを分類し，その際に生成される木々で「同じ葉ノードに属する回数が多いデータ同士の類似性は高い」という考えに基づき，データ間の類似度行列 (proximity matrix) を得る．すなわち，各サンプル同士が T 個の決定木の中で同じ葉ノードに現れた回数の平均を類似度とみなす．

距離尺度として一般的なユークリッド距離は変数間の相関を考慮できないが，RF によって算出される類似度では，変数間の関係を考慮した類似度を求めることができる．これは，RF が，変数間の統計的関連性を木モデルの中に取り込むことができることに加え，ランダムに変数を選択して算出した結果を混合することで精度の高い推測を可能とするためである．

サンプル間の類似度が算出されれば，様々な統計解析手法に用いることが可能である．例えば，多次元尺度構成法（MDS：Multi Dimensional Scaling）などを適用することも可能であるが，本論文ではデータ間の類似性をもとにデータの集合をグループ分けするクラスタリングを考える．

3. 提案手法

3.1 概　　要

従来は欠損を含むデータのクラスタリングのために類似度を導出する際に，前処理として平均値代入などの欠損値補完を行う必要があった．しかし，欠損値補完では類似度算出のために擬似的な完全データを生成できる一方で，観測されたデータと補完データを同等に扱って類似度を算出してしまうため，クラスタリング精度の低下をもたらす可能性がある．

そこで本研究では，欠損値を補完するのではなく，ランダムに抽出した（欠損データを含まない）部分完全データを活用する方法を考える．具体的には，欠損を含む元データから欠損を含まないデータの部分集合である縮退行列をランダムに複数生成し，各縮退行列から RF による類似度行列を算出する方法を提案する．一般に，欠損を含まない完全データから推定した類似度の精度は，欠損データを含むデータから推定した類似度よりも高いため，これらのアンサンブルによって，より精度の高い類似度行列が得られると期待できる．

一方，クラスタリングのための手法として，クラスタアンサンブル [13] が提案されているが，本研究の提案法は，欠損を含むデータから類似度行列を得るためにランダムな変数の抽出をしているという点で異なる．

3.2 RF を用いた統合類似度行列の算出

i 番目のサンプル $\boldsymbol{x}_i$（サンプル i とよぶ）の j 番目の変数の値 $x_{i,j}$（$1 \leq i \leq N, 1 \leq j \leq M$）を要素としてもつ，欠損値を含む $N \times M$ のデータ行列 $\boldsymbol{D} = [x_{i,j}]$ が与えられたもとで，サンプル間の類似度を求める問題を考える．ここで，$x_{i,j} \in \mathcal{R} \cup \{\phi\}$ であり，$\mathcal{R}$ は実数全体の集合，ϕ はデータの欠損を表すものとする．

3.2.1 縮退行列の生成

データ行列 $\boldsymbol{D}$ において全 M 個の変数の中から Q 個の変数（$Q < M$）の組をランダムに選択し，それらの変数全てに関して欠損値のないサンプルのみから成るデータ行列を生成する．このデータ行列を縮退行列とよぶ．上記の操作を K 回繰り返し，k 回目の変数選択における縮退行列を $\boldsymbol{D}_k = [\tilde{x}^k_{i,q}] \in \mathcal{R}^{N_k \times Q}$ $(1 \leq k \leq K, 1 \leq i \leq N_k, 1 \leq q \leq Q)$ と表す．ここで，$\tilde{x}^k_{i,q}$ はサンプル i の縮退後の q 番目の変数の値を表し，N_k は $\boldsymbol{D}_k$ に含まれるサンプル数を表す $(N_k \leq N)$．元のデータが欠損を含む $N \times M$ 行列であったのに対し，$\boldsymbol{D}_k$ は欠損を含まない完全な $N_k \times Q$ 行列となる．

3.2.2 類似度行列の統合

縮退行列 $\boldsymbol{D}_k$ から RF により類似度行列を生成し，これを $\boldsymbol{S}_k = [s^k_{i,i'}]$ とする．類似度行列 $\boldsymbol{S}_k$ は $N_k \times N_k$ の対称行列であり，縮退行列 $\boldsymbol{D}_k$ のもとで算出されたサンプル i とサンプル i' 間の類似度 $s^k_{i,i'} \in [0,1]$ を要素にもつ．K 個の類似度行列 $\boldsymbol{S}_k$ を最終的に一つの行列に統合し，これを統合類似度行列とよぶ．統合類似度行列は $N \times N$ の対称行列 $\boldsymbol{T} = [t_{i,i'}]$（$t_{i,i'} \in [0,1] \cup \{\phi\}$）で表される．$\boldsymbol{T}$ の要素 $t_{i,i'}$ は，各 $\boldsymbol{S}_k$ におけるサンプル i と i' の類似度の総和を，K 個の $\boldsymbol{S}_k$ のうち i と i' が共起した回数で割ることにより平均化して求める．K 個のいずれの $\boldsymbol{S}_k$ においても i と i' が共起していな

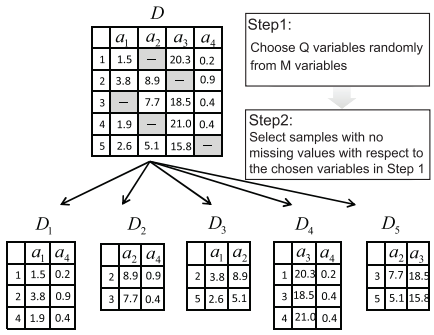

図 1　縮退行列の生成
Fig. 1　Generation of the degeneration matrix.

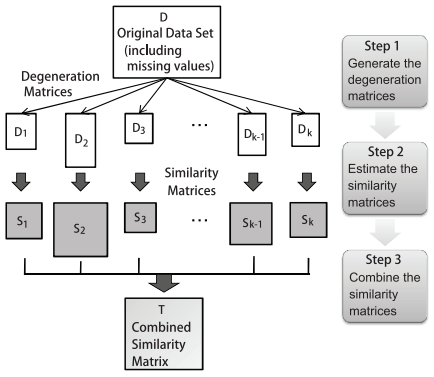

図 2　統合類似度行列の生成
Fig. 2　Estimation of similarity matrix.

い場合には $t_{i,i'} = \phi$ となる．この場合には，元の行列 $\boldsymbol{D}$ を平均値推定で補完し，これに対して RF から生成した類似度行列による類似度で代用する．

3.3　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．

Step1)　データ行列 $\boldsymbol{D}$ から縮退行列 $\boldsymbol{D}_k$（$1 \leq k \leq K$）を生成する．

Step2)　Step1 で生成した各縮退行列 $\boldsymbol{D}_k$ に対して RF の教師なし学習を適用し，類似度行列 $\boldsymbol{S}_k$（$1 \leq k \leq K$）を得る．

Step3)　得られた K 個の $N_k \times N_k$ の類似度行列 $\boldsymbol{S}_k$ を，$N \times N$ の一つの統合類似度行列 $\boldsymbol{T}$ に統合する．

4.　実験と結果

4.1　実 験 条 件

提案手法の有効性を示すため，以下の 2 通りの実験を行った．完全データの一部を欠損させたデータに対して類似度行列（ここでは距離も類似度と呼ぶ）を算出し，完全データから求めた類似度との平均 2 乗誤差を比較する（実験 1）．更に，これらの類似度をクラスタリングに適用した場合の精度の比較を行う（実験 2）．実験には，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 $N = 178$，次元数は $M = 13$，カテゴリー数は $C = 3$ 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，繰り返し回数は $K = 100$ とした．RF における決定木の生成法は CART [12] を用い，木の数は 500，分岐の際に選択する変数は 2 とした．また実験 2 では，ウォード法による階層クラスタリングを用いた(注1)．両実験共に，欠損データとするために元の完全データをランダムに 10〜60%欠損させる．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．

提案手法（Proposed）に対する比較手法としては，

(1)　平均値推定により欠損値補完したデータに対するユークリッド距離（MEAN+EUC）

(2)　平均値推定により欠損値補完したデータに対する RF の類似度（MEAN+RF）

(3)　縮退行列から求めたユークリッド距離を統合する手法（DEG+EUC）

とした．また，

(4)　Chained Equations による多重代入法 [10] で欠損値補完を行ってからクラスタリングする方法（MICE）

とも比較した．

4.2　評 価 方 法

実験 1 では，元の完全データから求めた類似度と，欠損データから算出された類似度との差異を，平均 2 乗誤差 MSE で評価した．元の完全データ $\boldsymbol{D}$ から求めた類似度行列を $\boldsymbol{Y} = [y_{i,i'}] \in \mathcal{R}^{N \times N}$，また，欠損データから導出した類似度行列を $\boldsymbol{R} = [r_{i,i'}] \in \mathcal{R}^{N \times N}$ とすると，MSE は式 (1) で表される．

$$MSE = \frac{\sum_{i=1}^{N} \sum_{i'=1}^{N} (r_{i,i'} - y_{i,i'})^2}{N \times N} \qquad (1)$$

実験 2 では，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と純度を用いた [14]．いま，L をクラスタ数，$\mathcal{C} = \{C_1, C_2, ..., C_L\}$ 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

(注1)：ウォード法ではサンプル間の距離（非類似度）行列を用いる必要がある．そこで RF によって得られた [0,1] の値をとる類似度行列に対しては，1 からこれらの値を引くことで距離行列に変換した．

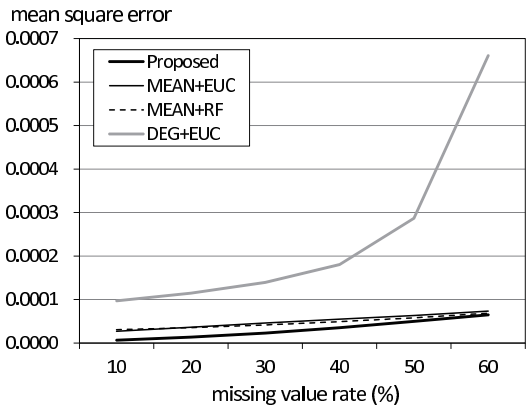

図 3　完全データから求めた類似度行列との平均 2 乗誤差（実験 1）
Fig. 3　Mean square error of similarity matrix.

果とし，また x_{ij} を C_i と A_j に同時に含まれるデータの個数とする．このとき，エントロピー E と純度 P はそれぞれ式 (2) と式 (3) で定義される．

$$E = \sum_{i=1}^{L} \frac{\left(\sum_{j=1}^{L} x_{ij}\right)}{N} E_i \tag{2}$$

$$P = \frac{1}{N} \sum_{i=1}^{L} \max_{h} |C_i \cap A_h| \tag{3}$$

ただし，$|\cdot|$ は集合の要素数を表す．また，エントロピーの式における E_i は

$$E_i = -\sum_{k=1}^{L} \frac{x_{ik}}{\left(\sum_{j=1}^{L} x_{ij}\right)} \log_{\mathrm{e}} \frac{x_{ik}}{\left(\sum_{j=1}^{L} x_{ij}\right)}$$

で与えられる．E，P の値はいずれも 0 から 1 の間をとり，E では値が低いほど，P では値が高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 1 の結果を図 3 に，実験 2 の結果を図 4（エントロピー）及び図 5（純度）に示す．

また，完全データから RF やユークリッド距離によって求めた類似度行列を用いてクラスタリングした場合のエントロピー，純度の値は以下の表 1 のとおりとなった．これは欠損率が 0%の場合に該当する．

実験 1 の結果（図 3）より，各欠損率において，提案手法による統合類似度行列の誤差が最小であることがわかる．したがって，アンサンブルを取り入れた提案手法では，欠損があっても完全データによる類似度に近い類似度を算出できることが示された．

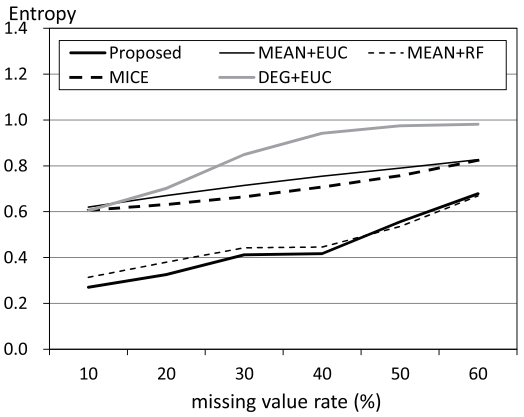

図 4　クラスタリングのエントロピー（実験 2）
Fig. 4　Entropy of clustering.

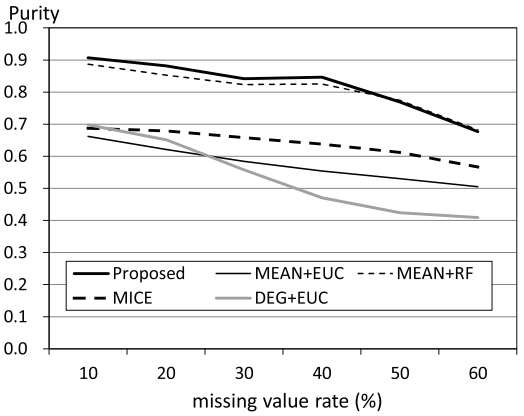

図 5　クラスタリングの純度（実験 2）
Fig. 5　Purity of clustering.

表 1　完全データから求めた類似度行列によるクラスタリング結果
Table 1　Result of clustering from complete data.

	RF にによる類似度	ユークリッド距離
エントロピー	0.203	0.607
純度	0.938	0.708

実験 2 の結果（図 4，図 5）より，類似度として RF による類似度行列を用いた手法（Proposed 及び MEAN+RF）がユークリッド距離を用いた手法（MEAN+EUC 及び DEG+EUC）や MICE よりも精度の高いクラスタリング結果となっている．このことから，クラスタリングの類似度として RF による類似度行列を用いることの有効性が示された．

また，欠損率 50%以上の部分では提案手法と MEAN+RF とのクラスタリング精度の差がほぼ見られなくなってしまうものの，欠損率 10〜40%にお

いては提案手法がすぐれている．また，提案手法とMICEとの比較では全ての欠損率で提案手法が高いクラスタリング精度を示している．この結果から，推定により欠損値を補完してから類似度行列を求める手法よりも，提案手法のように縮退行列から多数の類似度行列を生成し，それらをアンサンブルする手法がクラスタリングに対して有効であることも示された．提案手法と MEAN+RF が 50%以上の欠損率でほぼ同等の精度となったのは，提案手法において，欠損値推定ができなかったデータ間の類似度には，平均値補完後に算出した RF の類似度を用いているためであると考えられる．この点の改良は今後の課題とする．

5. むすび

本研究では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として，縮対行列に対し RF を適用する新たなアルゴリズムを提案した．数値実験より，提案した類似度の有効性，並びにそれをクラスタリングに適用した際の有効性が確認された．

なお，RF により作成された類似度行列は，データ間の類似度尺度の一つであると考えることができるが，欠損値が補完されたデータ同士の類似度（若しくは非類似度）を計る尺度には様々な手法が存在しているため，どのような尺度がクラスタリングに適しているかについては更に検証する必要がある．また，今後の課題として，欠損率を考慮した変数選択や欠損率と最良の変数選択数との関係の考察などが挙げられる．

謝辞　本研究にあたり，手厚いサポートを頂いた早稲田大学経営システム工学科の荒川貴紀氏，早川真央氏に深く感謝致します．本研究の一部は，科学研究費 (23510192) の助成を受けたものである．

文　献

[1] L. Breiman, “Random forests,” Machine Learning, vol.45, pp.5–32, 2001.

[2] 石岡恒憲，“Random Forest を用いた欠測データの補完に基づく大学入試センター試験科目間得点差,” 応用統計学，vol.40, no.3, pp.193–210, 2011.

[3] C. Englund and A. Verikas, “A novel approach to estimate proximity in a random forest: An exploratory study,” Expert Systems with Applications: An International Journal, vol.39, no.17, pp.13046–13050, Dec. 2012.

[4] A. Liaw and M. Wiener, “Classification and regression by randomForest,” R News, vol.2/3, pp.18–22, 2002.

[5] T. Shi and S. Horvath, “Unsupervised learning with random forest predictors,” Journal of Computation and Graphical Statistics, vol.15, no.1, pp.118–138, 2006.

[6] C.K. Enders, Applied Missing Data Analysis, Guilford, New York, 2010.

[7] A.C. Acock, “Working with missing values,” Journal of Marriage and Family, vol.67, no.4, pp.1012–1028, Nov. 2005.

[8] M. Huisman, “Imputation of missing item responses: Some simple techniques,” Quality & Quantity, vol.34, no.4, pp.331–351, Nov. 2000.

[9] S. van Buuren, Flexible Imputation of Missing Data, Chapman and Hall/CRC, 2012.

[10] K. Oudshoorn, S. van Buuren, and J. van Rijckevorsel, “Flexible multiple imputation by chained equations of the AVO-95 survey,” TNO report PG/VGZ/99.045, 1999.

[11] S.M. Iacus and G. Porro, “Missing data imputation, matching and other applications of random recursive prttitioning,” Computational Statistics & Data Analysis, vol.52, pp.773–789, 2007.

[12] T.S. Lim, W.Y. Loh, and Y.S. Shih, “Comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms,” Machine Learning, vol.40, pp.203–228, 2000.

[13] A. Strehl and J. Ghosh, “Cluster ensembles - A knowledge reuse framework for combining multiple partitions,” The Journal of Machine Learning Research, vol.3, pp.583–617, 2002.

[14] 新納浩幸，R で学ぶクラスタ解析，オーム社，2007.

（平成 25 年 6 月 30 日受付，9 月 5 日再受付）