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

1X09C046-4　真田祐希
指導教員　後藤正幸

1　研究背景と目的

データマイニング手法の一つにクラスタリングがある．クラスタリングでは，ユークリッド距離などの尺度を用いてサンプル間の類似度を算出し，類似性の高いサンプル同士を同一のクラスタとして併合する．データに欠損がなければ類似度の算出が可能である一方，データに欠損が含まれる場合には直接類似度を計算することが不可能となる．しかし，顧客の購買履歴データといった実データでは未購買アイテムなどが含まれるため，欠損を含むことが多く，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して類似度を計算する方法 [1] であるが，この方法では欠損率が高くなるにつれて多くの補完データが用いられるため，必ずしも適切な類似度が算出されるとは限らない．

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，欠損を含まない部分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，全てのデータ間の類似度が得られる保証がないため，何らかの工夫が必要である．

そこで本研究では，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方法を提案する．類似度の算出には，Random Forest（以下，RF）を用いた類似度行列 [2] を用いる．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手法による学習器として注目されており，ランダムに抽出した完全データの学習結果を統合する本手法との相性が良いと考えられる．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う．

2　準備

2.1　Random Forest

Random Forest [2] は複数の決定木 [3] を用いたアンサンブル学習アルゴリズムであり，高速かつ高精度で分類や回帰，クラスタリングを行えることから，様々な分野で利用されている．RF 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．

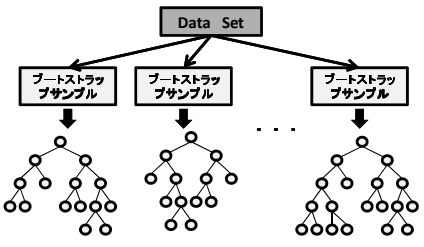


図 1．Random Forest

2.2　RF による類似度行列

類似度行列は，二つのサンプル間の類似度を要素に持つ対称行列である．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．

RF によりサンプル同士の類似度行列を生成できるが [2]，これ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つ．RF によるサンプル間の類似度は，生成された個々の決定木においてサンプル同士が同じ葉ノードに属した回数に基づいて算出される．この類似度行列をクラスタリングの距離尺度に適用することで，より正確なクラスタリングが可能になると考えられる．

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なう必要があった．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．算出された全ての類似度行列を統合することで統合類似度行列を得る．

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（サンプル i とよぶ）の j 番目の変数の値 $x_{i,j}(1 \leq i \leq N, 1 \leq j \leq M)$ を要素として持つ，欠損値を含むデータ行列 $\boldsymbol{D} = [x_{i,j}] \in \mathcal{R}^{N \times M}$ が与えられたもとでサンプルをクラスタリングする問題を考える．以下で，欠損を含む $\boldsymbol{D}$ から類似度行列を生成する方法について示す．

3.2.1　縮退行列の生成

データ行列 $\boldsymbol{D}$ において M 個の変数から 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データ行列を生成する．これらのサンプルから成る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 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 $\boldsymbol{S}_k = [s^k_{i,i'}] \in \mathcal{R}^{N_k \times N_k}$ とする．類似度行列 $\boldsymbol{S}_k$ は，縮対行列 $\boldsymbol{D}_k$ 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 $s^k_{i,i'}$ を要素に持つ行列である．K 個の $\boldsymbol{S}_k$ を最終的に一つに統合することで得られる $N \times N$ の統合類似度行列を $\boldsymbol{T} = [t_{i,i'}] \in \mathcal{R}^{N \times N}$ と定義する．$\boldsymbol{T}$ 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．

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 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．

Step4) 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' に関しては，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を代入する． □

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，完全データから算出される類似度（ここでは距離も類似度とよぶこととする）をあるべき姿の類似度とし，一部を欠損させたデータから算出した類似度との平均二乗誤差を比較する（実験 1）．さらに，これらをクラスタリングに適用した場合の精度の比較を行なう（実験 2）．

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 $N = 178$，次元数は $M = 13$，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，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 [3] 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．両実験共に，欠損データとするために元のデータを 10〜60 %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．比較手法としては，平均値推定による欠損値補完したデータに対するユークリッド距離（EUC+mean）と RF の類似度（RF+mean）とした．

4.2 評価方法

実験 1 では，元の完全データから算出された類似度と，欠損データから算出された類似度との差異を，平均二乗誤差 MSE で評価した．元の完全データ $\boldsymbol{D}$ から導出した真の類似度行列を $\boldsymbol{Y} = [y_{i,i'}] \in \mathcal{R}^{N \times N}$，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．

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

MSE の値が小さい程，欠損の影響が少ない類似度が得られていることを示す．

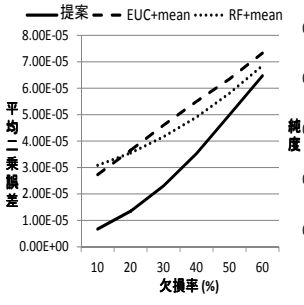
実験 2 では，クラスタリングの性能の評価方法として一般的に用いられる指標である純度を用いた．純度 P は L をクラスタ数，$\mathcal{C} = \{C_1, C_2, ..., C_L\}$ をクラスタリング結果，$\mathcal{A} = \{A_1, A_2, ..., A_L\}$ を正解となるクラスタリング結果としたとき，式 (2) で定義される．

$$P = \frac{1}{N} \sum_{i=1}^{L} \max_{h} | C_i \cap A_h | \quad (2)$$

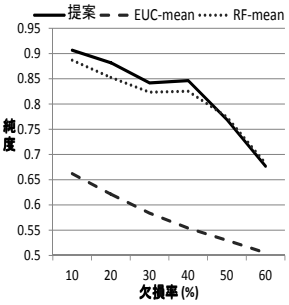
P の値は 0 から 1 の間をとり，値が高いほどクラスタリング結果が良好であることを意味する．ただし，$|\cdot|$ は集合の要素数を表す．

4.3 結果と考察

実験 1 の結果を図 2 に，実験 2 の結果を図 3 に結果を示す．



(左) 図 2. 真の類似度との平均二乗誤差 (実験 1)



(右) 図 3. クラスタリングの純度 (実験 2)

実験 1 の結果（図 2）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．

実験 2 の結果（図 3）より，距離尺度として RF の類似度を用いた手法がユークリッド距離を用いた手法を大きく上回るクラスタリング結果となっている．このことから，クラスタリングの距離尺度として RF の類似度を用いることの有効性が示された．また，欠損率 10 %〜40 %において，提案手法は欠損値補完後に RF の類似度を算出する手法より優れている．したがって，クラスタリングに提案手法による類似度を用いることの有効性も示された．50 %以上の欠損率では 2 手法の差はほぼ見られないが，これは提案手法において，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 の類似度を代入しているためであると考えられる．この点の改良については今後の課題とする．

以上のことから，提案した統合類似度行列の有効性，並びに提案した統合類似度行列をクラスタリングに用いることの有効性が示された．

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，提案した統合類似度行列の有効性，並びにそれをクラスタリングに適用することの有効性が確認された．今後の課題としては，欠損率を考慮した変数選択，欠損率と最良の変数選択の数 Q の関係性の考察などが挙げられる．

参考文献

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

[2] LEO BREIMAN, "Random Forests," *Machine Learning, 45*, pp.5-32, 2001.

[3] Tjen-Sien Lim, Wei-Yin Loh, Yu-Shan Shih, "Comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms," *Machine Learning,40*, pp.203-228, 2000.