ISSN 2186-5647

P-16

自己組織化マップを用いた欠損データの推定

日大生産工（学部）　○杉浦　宏太郎　日大生産工　山内　ゆかり

1. まえがき

近年、情報技術の発達によって大規模データを解析して新たな知識を発見し、有効活用しようとする試みであるデータマイニングの重要性が増加している。

しかし、実際に収集されるデータは、様々な理由から、一部の値が欠損していることがあり、このような欠損データへの対処法は、データマイニングにおける問題の一つとなっている。

欠損データへの対処法には、欠損を含むサンプルや変数を削除する手法や欠損している値を推定して補完する手法[1]があり、様々な研究が行われている。この問題に対して、菊池らは、欠損のあるデータに対し、複数の自己組織化マップを用いてのデータの補完手法を提唱した[2]。これにより、上記の問題を改善し、欠損割合の大きい非線形な規則性を持ったデータに対しても、適切な推定や補完を可能とすることを試みた。しかし、欠損データの補完に際し、欠損データの実際の値と推定値の誤差が学習時の誤差として累積してしまう問題や欠損部ごとにマップを生成しなければならないため、欠損部によってはマップ数が膨大な数になってしまう問題がある。

そこで、本研究では、マップ数増大の問題の解決と誤差の減少化を目的とする。具体的には、欠損を持つ全データを単一の自己組織化マップを用いて学習させ、学習後のマップから欠損部の推定値を得る方法を提案する。

2. 従来研究

菊池らは、米国カリフォルニア大学アーバイン校のサイトで公開されている学習用データセット中の、アヤメのデータ[3]を元に作成したサンプルを学習に用いた。アヤメのデータはサンプル数が150個で、要素を四つ持っている。要素はそれぞれ「萼片の長さ」「萼片の幅」「花弁の長さ」「花弁の幅」を表しAttribute1~4と定義する。136個のサンプルを要素ごとに欠損させたサンプルをGroup1からGroup4の四つに分けて(一つのGroup毎に34個)作成し、Group0を14個の欠損のないサンプル群とし、テスト用データとする。以下の表1にGroup分け概念図を示す。

学習のマップは要素数分の四つ作成し、それぞれのGroupの番号に対応させるMAPを作成する。各MAPではGroup0のサンプル群と対応するGroupのサンプル群を用いて学習を行う。以下の図1に学習の概念図を示す。

表1　Group分け概念図

	Attribute1	Attribute2	Attribute3	Attribute4
Group0のサンプル(14個)	欠損なし	欠損なし	欠損なし	欠損なし
Group1のサンプル(34個)	欠損	欠損なし	欠損なし	欠損なし
Group2のサンプル(34個)	欠損なし	欠損	欠損なし	欠損なし
Group3のサンプル(34個)	欠損なし	欠損なし	欠損	欠損なし
Group4のサンプル(34個)	欠損なし	欠損なし	欠損なし	欠損

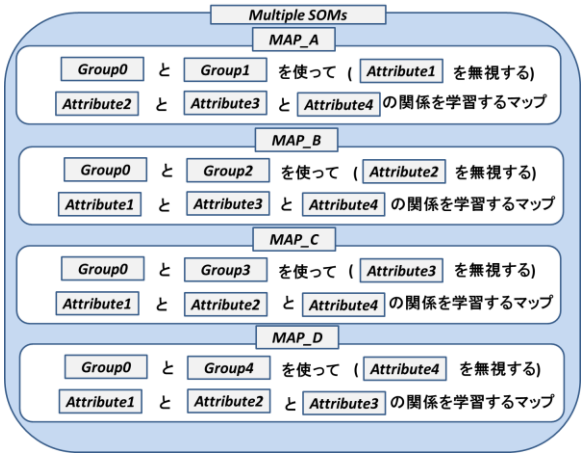

図1　学習の概念図

学習後の各マップの全ノードと欠損のあるサンプルを比較し、各マップからサンプルと最も近似しているノードを以下の式(1)を用いて決定する。

$$|x(t) - m_c(t)| = min\{|x(t) - m_i(t)|\} \quad (1)$$

$x(t)$は学習時間tの入力サンプルのベクトル、mは学習時間tのマップのノードのベクトル表している。ユークリッド距離$|x(t) - m_i(t)|$を最小にするノードm_iを探して、それに添え字cをつけ、配列に格納する。

その後、決定されたノードの各要素の平均

An Estimation of Missing Data Using Self-Organizing Map
Kotaro SUGIURA and Yukari YAMAUCHI

をそのサンプルの欠損部の推定値（E）とし、欠損部の補完を行う。補完された値と真値との誤差の検定には以下の式(2)を用いて行う。

$$e_{ij} = \frac{|E_{ij} - O_{ij}|}{\sqrt{V_j}} \qquad (2)$$

サンプル群において、i番目のサンプルのj番目の要素に欠損があるとする。その要素の推定値をE_{ij}、真値をO_{ij}とする。このとき、E_{ij}が良い推定値かどうかの判断に、相対誤差e_{ij}を用いる。ここでV_jは、元のサンプル群におけるi番目の要素の分散を表す。V_jは以下の式(3)を用いて求める。

$$V_j = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad (3)$$

多変数データでは要素によって単位が異なるため、異なる要素間での誤差の比較を可能とするために、このようにして無次元化を行う。その後、さらに補完を施したサンプルに含まれる全欠損に対する、e_{ij}の平均値e_{ave}を計算した。以下の式(4)に平均値計算を示す。

$$e_{ave} = \frac{1}{\sqrt{ij}}\sum_{j=0}^{m}\sum_{i=0}^{n} e_{ij} \qquad (4)$$

e_{ave}の値が小さいものが誤差が少ない。

3. 提案手法

本研究では、学習用マップの増大の問題の解決と推定値と真値の誤差の減少化の実現のために、単一の自己組織化マップを用いて欠損部を持つサンプル全てを学習させて推定値を得る方法を提案する。

下記に提案手法のアルゴリズムを示す

1. 欠損を持つ全サンプルを単一の自己組織化マップで学習する。
2. 学習後のマップから、学習後のマップの全ノードと欠損のあるサンプルを式(1)を用いて比較し、マップからサンプルと近似しているノードを決定する。
3. 決定されたノードの要素を、そのサンプルの欠損部の推定値（E）とし、推定値と真値の誤差の検証および推定値の妥当性の検証を上記の式(2)、式(3)、式(4)を用いて行う。

4. 実験環境

実験環境の初期画面を以下の図 2 に示す。マップのノード数を、二次元配列を用いて 100 個として設定し、学習回数を 1500 回と設定し、学習を行った。学習後のマップを以下の図 3 に示す。

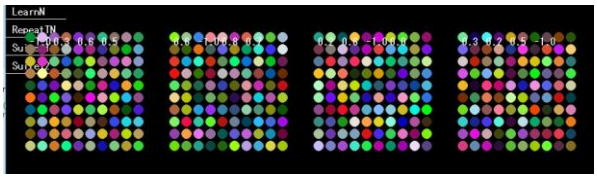

図 2 初期画面

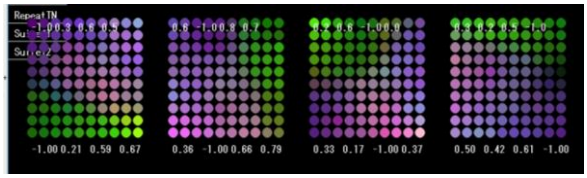

図 3　学習後結果

図 3 が学習後のマップの状態である。現在、従来手法の複数のマップを用いての学習と近似ノードの決定を完了している。各類似ノードの値は以下の表 2 に示す。-1.00 は欠損部を表している。

表 2 各マップの類似ノードの値

MAP_A	-1.00	0.21	0.59	0.67
MAP_B	0.36	-1.00	0.66	0.79
MAP_C	0.33	0.17	-1.00	0.37
MAP_D	0.50	0.42	0.61	-1.00

5. まとめ

本研究では、学習用マップの増大の問題の解決と推定値と真値の誤差の減少化の実現のために、単一の自己組織化マップを用いて欠損部を持つサンプル全てを学習させて推定値を得る方法を提案した。提案手法の今後の課題は、決定された類似ノードから推定値を求め、推定値と真値の誤差の検証および推定値の妥当性の検証を行うことである。

参考文献

1) 和久屋寛, 永野俊, 部分データ対応型自己組織化マップにおける正答率改善の試み, 26th Fuzzy System Symposium（Hiroshima, September13-15, 20）

2) 菊池悠意, 岡田伸廣, 辻康孝, 複数自己組織化マップを用いた欠損データの推定, 日本機械学会論文集（c 編）79 巻 806 号(2013-10)

3) University of California, Irvine, “Iris Data Set”, Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/Iris (参照日　2017 年 10 月 13 日).