ニューラルネットワークによる 賃貸物件データセットの欠損値補完の一手法

加藤　暢之† 新妻弘崇†† 太田　学††

†, †† 岡山大学大学院自然科学研究科 〒700–8530 岡山県岡山市北区津島中三丁目 1 番 1 号
E-mail: †kato@de.cs.okayama-u.ac.jp, ††{niitsuma, ohta}@cs.okayama-u.ac.jp

あらまし 機械学習やニューラルネットワークの多くは完全なデータを用いてモデルを訓練する．しかし，実際に存在するデータのほとんどは欠損値を含んでいるため，欠損値補完処理を行う必要がある．そこで本稿ではニューラルネットワークを用いた欠損値補完の一手法を提案する．手法の評価には LIFULL HOME'S データセットの一部を用い，その欠損率は 36.75%である．このデータセットに対して，平均値補完，主成分分析による補完などと比較して提案手法の欠損値補完の精度を評価する．

キーワード 賃貸物件情報，ニューラルネットワーク，欠損値補完

1 はじめに

データ中の欠損値を扱うアプローチには，欠損値を含むデータの除去，平均値などによる単純な補完，多重代入法などが存在する [1]．一般的には欠損率が低い場合には欠損値の除去，欠損率が高い場合にはランダムな値や平均値による補完が用いられることが多い．中間層を深くした深層学習において ReLU をドロップアウトと共に用いると，ドロップアウトによるランダム欠損によって欠損値の補完に相当する処理が自動的にできてしまうことがある．そのため近年は，この深層学習の性質から欠損値補完に重点を置かないニューラルネットワークが利用されることも多い．

しかし，深層学習において中間層を増やすことは学習パラメータの指数関数的な増加に繋がるため，中間層を重ねると計算コストが急激に増加する．膨大なパラメータを学習できる処理性能と時間がない場合や，モデル自体が巨大な場合には，欠損値をランダムに補完して中間層で吸収する手法はコストが高い．

また，深層学習の中間層では一部のモデルを除いて可読性が低いため，一般的にはモデルが入力データからどのような特徴を学習しているか確認することは困難である．一方，オートエンコーダは出力を入力データに近付けることを目的としているため，入力データと出力の比較によりモデルがどのような特徴を学習したか，ある程度判別できる．データの特徴を知ることはモデルの構築において重要であるため，獲得した特徴を確認できることはオートエンコーダによる補完のメリットでもある．

さらに転移学習によってオートエンコーダのモデルと学習した重みをそのままに，モデルの層をを追加することもできる．これにより深層学習においても事前学習により計算量の削減とより可読性の高いモデルを作成できる可能性がある．

これらを踏まえて本稿では，モデルのチューニングコストの小さいシンプルなオートエンコーダによる欠損値補完の一手法を提案する．

第 2 節で欠損値補完に関連した研究について述べ，第 3 節で提案したモデルについて述べる．また，第 4 節で提案手法について比較評価実験を行い，第 5 節でその考察をする．最後に第 6 節でまとめを述べる．

2 関連研究

2.1 主成分分析

主成分分析による欠損値補完は多重代入法の代表的な手法であり，S. Dray ら [2] によって欠損値補完における有効性が示されている．欠損値を含むデータに対する主成分分析は欠損値を無視するものと欠損値を考慮したものに分けられる．欠損値を無視する手法には GowPCoA [3] や PairCor，欠損値を考慮した手法には NIPALS [4]，Ipca [5] などが存在する．本稿では欠損値を考慮した主成分分析手法である NIPALS による欠損値補完を比較対象の一つとした．

NIPALS [4] は本来は完全なデータセットに対する主成分スコアとローディングを求めるアルゴリズムである．しかし，欠損したエントリに対する重みを 0 とした重み付き回帰により容易に拡張できる．そのため拡張した NIPALS は欠損値を含むデータの主成分分析に使用される．NIPALS のアルゴリズムでは，まず第一次元の主成分スコア F_1，ローディング V_1 を交互最小二乗法 [6] により推定する．第二次元のスコア F_2，ローディング V_2 を残差行列 $X - F_1 V_1^T$ を用いて次元間の直交性を保持したまま次元を圧縮する．この処理を繰り返して主成分を求める．欠損値補完における NIPALS の有効性は S. Dray ら [7] によって示されている．

2.2 オートエンコーダ

オートエンコーダとは，データ特徴の獲得を目的としたニューラルネットワークである．オートエンコーダのモデルは入力データを潜在変数に変換するエンコーダと，潜在変数から入力

を復元するデコーダから構成される．このとき潜在変数の次元が入力より少ない場合は主成分分析と同様に次元削減を行うことになる．オートエンコーダでの損失関数は式 (1) で表される入力データと出力データの差である．この 2 つが近づくよう誤差逆伝播法でニューラルネットワークの重みを更新する．

$$L = \sum_{n=1}^{N} \|\boldsymbol{x_n} - \boldsymbol{y_n}\|^2 \tag{1}$$

式 (1) はオートエンコーダの再構成誤差 (Mean Absolute Percentage Error, MAPE) と呼ばれる．ここで $\boldsymbol{x_n}$ は n 番目の入力データ，$\boldsymbol{y_n}$ は n 番目の出力である．この再構成誤差は 4 節の評価実験の評価指標としても用いる．

2.3 スパースオートエンコーダ

スパースオートエンコーダはオートエンコーダの潜在変数にスパース性を持たせたオートエンコーダである．単純にオートエンコーダの潜在変数の次元を大きくすると，入力データのコピーを出力することになるが，損失関数にスパース項による制約を加えることで入力データの特徴を獲得する．スパースオートエンコーダの損失関数は式 (2) で定義される [8].

$$L = \sum_{n=1}^{N} \|\boldsymbol{x_n} - \boldsymbol{y_n}\|^2 + \alpha \sum_{d=1}^{D} KL(\rho \| \hat{\rho_d}) \tag{2}$$

式 (2) の第一項は式 (1) と同様に入力と出力の差であり，第二項はスパース項である．第二項の D は潜在変数の次元であり，α はスパース項の強さを調節するパラメータである．KL はカルバック・ライブラー情報量 [8], [9] である．$\hat{\rho_d}$ は d 番目の潜在変数の平均活性度であり，$\boldsymbol{x_n}$ が入力されたときの d 番目の潜在変数の値 $Z_d(\boldsymbol{x_n})$ により以下の式 (3) で表される．

$$\hat{\rho_d} = \frac{1}{N} \sum_{n=1}^{N} Z_d(\boldsymbol{x_n}) \tag{3}$$

ρ は平均活性度の目標値であり，ρ を小さくすることで，再構成誤差と潜在変数の平均活性度が小さくなるよう学習され，スパースな中間層を構築できる．ρ を大きくするとスパース性が弱くなり，より多いパターンに学習データの特徴を押し込めようとする．逆に ρ を極端に小さくするとスパース性が強くなり，入力データのそれぞれの特徴をより強く学習する．データによって学習すべき特徴の量は異なるため，ρ の大きさと潜在変数の次元数は入力とするデータによってチューニングする必要がある．

また，G. Lovedeep ら [10] の提案したモデルでは入力データと中間層をドロップアウトすることで再構成精度が向上した．

3 提 案 手 法

3.1 データセット

本稿で使用したデータセットは国立情報学研究所が提供している賃貸物件情報データセットである LIFULL HOME'S データセットである．LIFULL HOME'S データセットには 2015 年 9 月時点での賃貸物件スナップショットデータ，高精細度間取り画像データ，賃貸・売買物件月次データが含まれており，本研究では賃貸物件スナップショットデータを対象とする．賃貸物件スナップショットデータのうち，説明変数として使用できると考える変数を表 1 に示す．

本研究では量的変数の中から，物件の客観的な数値を表していると思われる 10 変数について欠損値の補完を行う．補完の対象とした量的変数とその欠損率を表 2 に示す．

全スナップショットデータは約 533 万件存在するが，岡山県内の賃貸物件データ 78,019 件を対象に実験を行う．また，岡山県内の賃貸物件データ 78,019 件のうち，上記の 10 変数全てについて欠損していないデータは 600 件のみである．そのため，欠損値を含むデータを除去する手法は有効ではないことが分かる．これらの賃貸物件データの状況から，本稿ではニューラルネットワークの訓練に欠損値を含む 77,419 件を使用し，オートエンコーダのモデルを訓練する．その後，欠損値を含まない 600 件のデータにランダムな欠損値を与え，元データの復元を行う．

本稿で使用する変数の例として，小学校までの距離のヒストグラムを図 1 に示す．

表 1　説明変数として使用できる変数

	変数名
量的変数	総戸数，空き物件数，バス停時間 1，バス停距離 1，バス停時間 2，バス停距離 2，建物面積，建物階数，築年数，部屋階数，間取り部屋数，賃料，共益費， 礼金，敷金，保証金，更新料，契約期間， 駐車場料金，駐車場距離，駐車場空き台数， 小学校距離，中学校距離，コンビニ距離， スーパー距離，総合病院距離
カテゴリ変数	物件種別，都道府県，市区郡町村，路線，駅 1，バス停 1，路線 2，駅 2，バス停 2，用途地域，都市計画，建物構造，新居/未入居，向き，間取り部屋種類，小学校名，中学校名

表 2　量的変数ごとの欠損率

量的変数	欠損率 (%)
最寄りのバス停 (sec)	82.0
二番目のバス停 (sec)	91.9
建物面積 (m^2)	0.1
築年数 (年)	0.1
駐車場距離 (m)	45.3
小学校距離 (m)	40.3
中学校距離 (m)	44.9
コンビニ距離 (m)	26.2
スーパー距離 (m)	29.2
総合病院距離 (m)	43.7

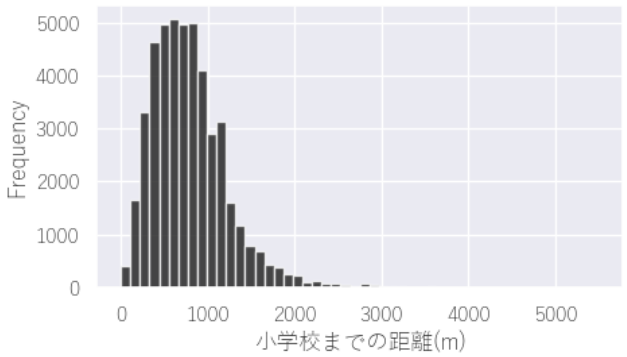

図 1　訓練データの小学校までの距離 (m) の分布

3.2 提 案 手 法

本稿ではスパースオートエンコーダをベースとしてドロップアウトにより疑似的にデータを欠損させたモデルを提案する．提案モデルでは，ドロップアウトによる欠損値を含むデータに対して欠損値によるノイズを抑えた元データの特徴の獲得が見込める．また，式 (2) の誤差関数に正則化項を加えた式 (4) により過学習を抑制している．式 (4) の第三項は L2 正則化項であり，実験では L2 正則化を用いているが，L1 正則化が適している場合も存在する．

$$L = \sum_{n=1}^{N} \|\boldsymbol{x_n} - \boldsymbol{y_n}\|^2 + \alpha \sum_{d=1}^{D} KL(\rho \| \hat{\rho_d}) - \beta(\|\boldsymbol{W}_e\|^2 + \|\boldsymbol{W}_d\|^2) \tag{4}$$

第一項，第二項は式 (2) と同様である．第三項について，β は正則化項の強さを調整するパラメータ，$\boldsymbol{W}_e$ はエンコーダの重み，$\boldsymbol{W}_d$ はデコーダの重みである．

本稿で提案するモデルは図 2 である．以下では各層について述べる．

Input Layer　選択した 10 次元の量的変数に対して，欠損値をランダムな値で補完したものを入力とする．

Dropout　入力データはランダムな値で欠損値を補完されているため，疑似的にデータの欠損を再現するために入力データの欠損率と等しい割合で入力をドロップアウトし Encoder へ渡す．

Encoder　入力と潜在変数の中間の次元数を持ち，Input Layer と Latent Variable の仲介をする．

Latent Variable　入力データの特徴を学習した潜在変数であり，本稿では入力データ次元の 10 倍の次元数に設定している．

Decoder　Encoder とは真逆の働きをする中間層であり，潜在変数から入力データへの再現を仲介する．

Output Layer　入力データから学習した特徴である潜在変数を元に入力データを再現する．Input Layer と Output Layer から復元誤差を算出し，誤差逆伝播法により重みの更新を行う．

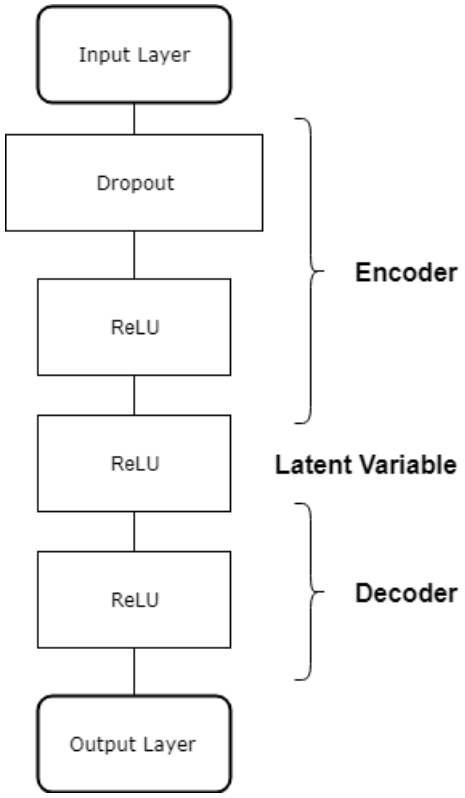

図 2　提案モデル

3.3 アルゴリズム

ここでは，3.2 節で述べた提案モデルを用いて欠損値を補完するアルゴリズムについて述べる．まず提案モデルのニューラルネットワークを以下の手順 **Train** で学習する．

Train

1. 入力データの欠損値をランダムな値で補完する．
2. 式 (2) の損失関数が収束するまでモデルを訓練する．
3. 入力データとモデルの出力から再構成誤差を算出する．
4. **1.** ∼ **3.** を n 回繰り返す．
5. 再構成誤差が収束していなければ **1.** に戻る．

次に手順 **Train** で学習したニューラルネットワークの性能を以下の手順 **Test** で評価する．

Test

1. テストデータの一部を欠損率に従い削除し，ランダムな値を代入する．
2. 訓練したモデルによってテストデータから出力を得る．
3. 欠損を与えた変数についてのみ再構成誤差を算出し，その総和をモデルの評価指標とする．

モデルの活性化関数は ReLU，最適化関数は RMSprop，正則化項には L2 ノルムを使用した．

4 評 価 実 験

本節ではランダム補完，平均値補完，主成分分析補完，提案手法のそれぞれについて，欠損率ごとに再構成誤差により比較する．ここでの再構成誤差は正規化したテストデータにおける補完の前後での絶対平均誤差を用いる．また，テストのため完

全なデータをランダムに欠損させるため，欠損値に依存した偏りが生じないよう，実験を 10 回行った平均を評価する．表 3 にそれぞれの欠損率における各手法の再構成誤差を示す．

欠損率が 0.2 と 0.5 のときに平均値補完の再構成誤差が最も小さい．これは使用したデータが正規分布に従っており，外れ値は存在しているが少数のため影響が少ないためと考えられる．そのため欠損率が 0.8 と大きくなってくると外れ値に対する補完ができない平均値補完は再構成誤差が極端に大きくなっている．

NIPALS による補完では平均値補完ほど欠損値の値の影響を受けていないが欠損率が 0.6 を超えると再構成誤差が大きく増加する．

提案手法による欠損値補完の結果のうち，小学校までの距離を補完した結果のヒストグラムを図 3 から図 12 に示す．提案手法では，欠損率 0.8 のときに再構成誤差が小さくなっているが，再構成した欠損値はほぼ単一値に定まっていた．

図 6 より，欠損率が低い場合には外れ値の影響を受けてはいるが，再構成前と類似した分布のデータが生成されていることが分かる．しかし欠損率が高くなると図 12 のように再構成後のデータがほぼ単一値に決まってしまう場合がほとんどであった．これはモデルを訓練するデータも欠損率が 0.8 になるため大部分のデータの重みが正規化項により小さくなり，データの特徴を学習できないためであると考えられる．欠損率が高いときに再構成誤差が小さくなっているのは頻度の最も高いデータの特徴のみをモデルが獲得しており，提案モデルでは欠損値が高い場合に，データの特徴を抽出できているといえる．

5 考　　察

本節では作成したモデルのパラメータチューニングと評価実験について考察する．なかでも，提案モデルでチューニングしたパラメータである正則化項 L1/L2，平均活性度の目標値，中間層のユニット数について述べる．

本稿で用いたデータは量的変数のみであるが L1 ノルムは正則化のペナルティが L2 ノルムに対して大きく，ほぼ全ての重みが 0 となった．一般的には平均活性度を大きく設定するとモデルの表現力が高まることが知られている．しかし，本稿で扱ったデータセットでは L1 ノルムを用いると必ず再構成後の値がほぼ単一値となった．このことから L1 正則化は量的変数の特徴を学習するには不向きである．

次に平均活性度の目標値について述べる．平均活性度は画像などのように対応する特徴が明確な場合には小さい値を設定す

表 3　手法ごとの再構成誤差 (MAPE)

補完手法	欠損率		
	0.2	0.5	0.8
ランダム補完	0.3684	0.3797	0.3694
平均値補完	0.1226	0.1325	0.3425
NIPALS 補完	0.1952	0.2067	0.2542
提案モデル補完	0.1861	0.1879	0.1747

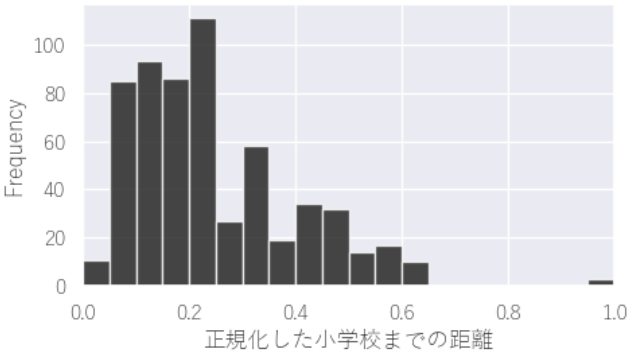

図 3　テストデータの小学校までの距離の分布

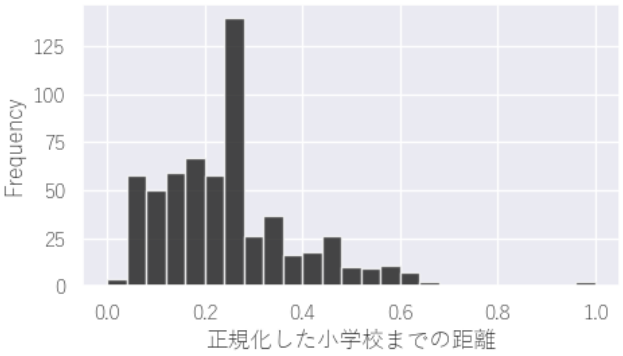

図 4　平均値補完した小学校までの距離の分布 (欠損率 0.2)

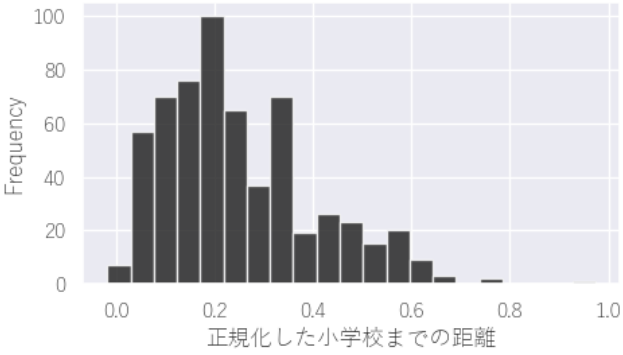

図 5　NIPALS で補完した小学校までの距離の分布 (欠損率 0.2)

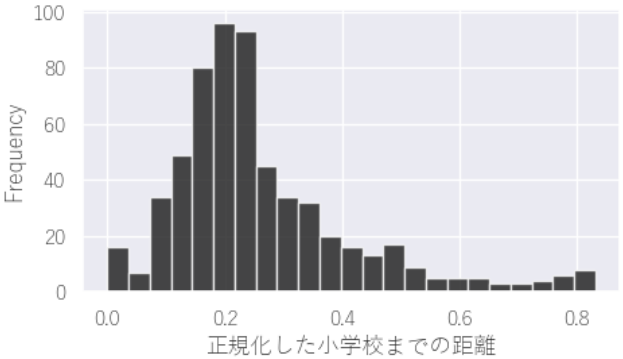

図 6　提案手法で復元した小学校までの距離の分布 (欠損率 0.2)

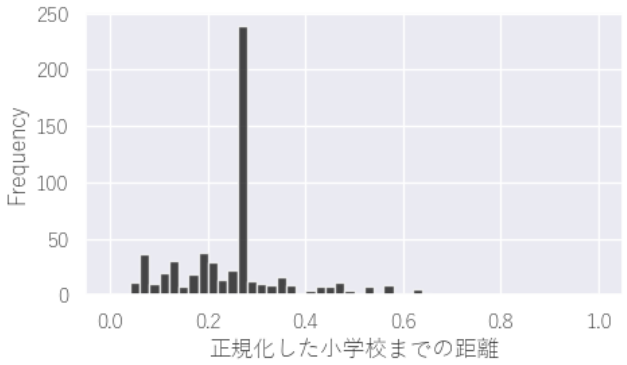

図 7　平均値補完した小学校までの距離の分布 (欠損率 0.5)

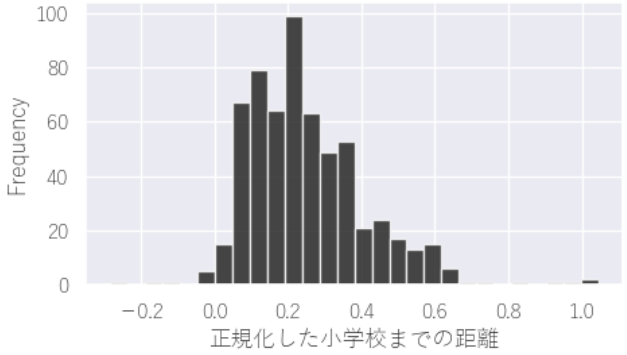

図 8　NIPALS で補完した小学校までの距離の分布 (欠損率 0.5)

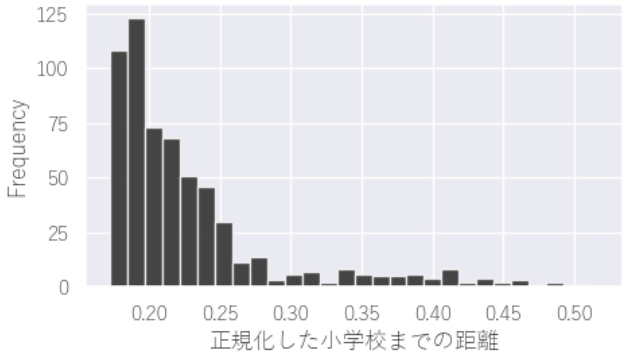

図 9　提案手法で復元した小学校までの距離の分布 (欠損率 0.5)

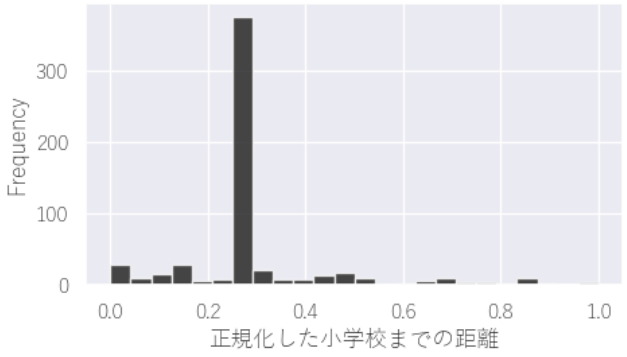

図 10　平均値補完した小学校までの距離の分布 (欠損率 0.8)

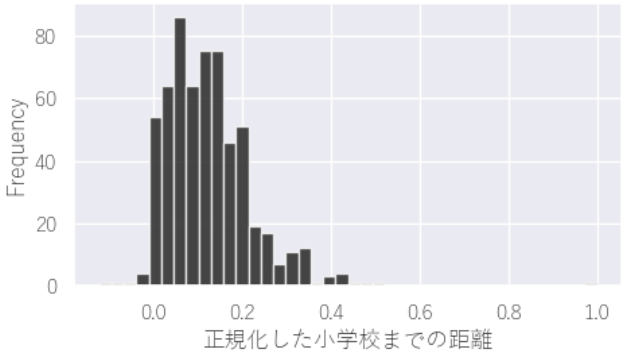

図 11　NIPALS で補完した小学校までの距離の分布 (欠損率 0.8)

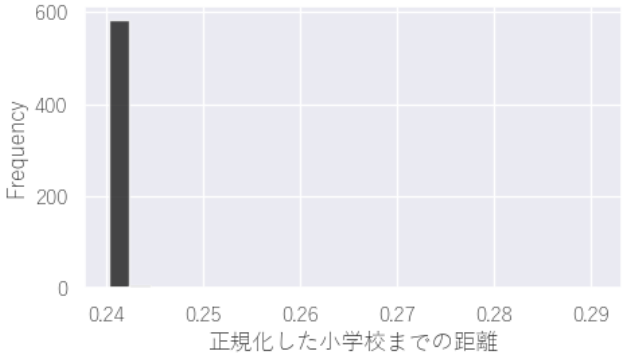

図 12　提案手法で復元した小学校までの距離の分布 (欠損率 0.8)

ることでより鮮明な特徴を学習できる．しかし本稿の賃貸物件データセットでは平均活性度を 0.01 のように小さく設定すると特徴を学習しなかった．このことから本稿で扱っている賃貸物件データセットには明確なパターンというものは存在していない可能性が高い．そのため平均活性度はモデルが特徴を学習し始めた 0.05 に設定した．ある程度の法則性が存在しているデータに対しては，平均活性度の目標値を小さく設定しても特徴を学習できる．

ニューラルネットワークにおいて入力次元数より大きな全結合層を加えることは入力の複製を出力することになる．そのため一般的には入力次元数より少ないユニットを持つ中間層を追加し，特徴の抽出を試みる．しかしスパースオートエンコーダでは正則化項により疎な中間層を構築するため，中間層のユニット数は入力次元数以上に設定する．平均活性度がデータの明確さを表現するのに対して，中間層のユニット数はデータの持つ特徴量を表現するものである．

6　ま　と　め

本稿ではスパースオートエンコーダを用いて欠損値を含むデータセットを使用してモデルを訓練し，データの特徴を学習することで欠損値を補完する手法を提案した．評価実験では平均値補完，主成分分析による欠損値補完手法である NIPALS との比較を行った．提案手法は欠損率に関わらず NIPALS より再

構成誤差が小さかったが，欠損率が低い場合には平均値補完より再構成誤差が大きくなった．

今後は本稿で提案したモデルの出力を入力とするニューラルネットワークモデルを構築し，賃貸物件データセットを活用する方法を検討する．

文　献

[1] J. W. Graham, A. E. Olchowski, T. D. Gilreath, “How many inputations are really needed? Some practical clarificatons of multiple imputation theory,” Prevention Science, Vol. 8, Issue 3, pp. 206–213, 2007.

[2] S. Dray, J. Josse, “Principal component analysis with missing values: a comparative survey of methods,” Plant Ecology, Vol. 216, Issue 5, pp. 657–667, 2015.

[3] J. C. Gower, “Statistical methods of comparing different multivariate analyses of the same data,” Mathematics in the Archaeological and Historical Sciences, pp.138–149, 1971.

[4] H. Wold, E. Lyttkens, “Nonlinear iterative partial least squares NIPALS) estimation procedures,” Bull Int Stat Inst 43, pp. 29-51.

[5] H. A. L. Kiers, “Weighted least squares fitting using ordinary least squares algorithms,” Psychometrika, Vol. 62, Issue 2, pp. 251–266, 1997.

[6] F. W. Young, Y. Takane, J. De Leeuw, “The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features,” Psychometrika, Vol. 43, Issue 2, pp. 279–281, 1978.

[7] S. Dray, N. Petorelli, D. Chessel, “Multivariate Analysis of Incomplete Mapped Data,” Transactions in GIS, Vol. 7, Issue 3, pp. 411–422, 2003.

[8] A. Ng, “Sparse autoencoder,” CS294A Lecture notes, p. 72, 2011.

[9] S. Kullback, R. A. Leibler, “On Information and Sufficiency,” Snn. Math Statist. 22, pp. 89–86, 1951.

[10] G. Lovedeep, Ke Wang, “MIDA: Multiple imputation using denoising autoencoders,” In Advances in Knowledge Discovery and Data Mining, pp. 260–272, 2018.