センサーデータのためのProgramming by Exampleに基づくデータ補完手法

永島 寛子[1]　加藤 由花[1]

概要：近年，基幹システムで扱うデータだけではなく，センサーデータやウェアラブルデバイスのデータなど，分析に利用可能なデータの量と種類が増えてきた．しかしながら，収集データを分析モデルに入力するためには，分析前に「前処理」を行う必要がある．前処理は，単位の統一や外れ値や欠損値の対処などを行う必要があり，分析者に多大なインパクトを与えている．そのため，分析者の前処理における負荷を削減し，かつ同程度以上の精度をもつ前処理の自動化が望まれている．本稿では，前処理で行われる作業のうち外れ値・欠損値をベイズ推論を用いて自動で補完する手法を提案する．自動補完では対処しきれない手法の一部を変更するなどのカスタマイズ性を持たせるため，Programming by Example アプローチを採用し，人の知識を例としてモデルに入力することにより人の知識と機械学習を融合させる．提案手法と既存手法による自動補完の補完精度を比較し，提案手法が既存手法より，センサーデータの補完手法として有効であることを検証した．

Data Imputation for Sensor Data based on Programming by Example

HIROKO NAGASHIMA[1]　YUKA KATO[1]

1. はじめに

近年，基幹システムで扱うデータだけではなく，センサーデータやウェアラブルデバイスのデータなど，分析に利用可能なデータの量と種類が増えてきた．多種多様なデータを活用した分析の例として，ショッピングモールの顧客動向の分析やロボットの自律行動，クレジットカードの不正検知などがある．しかしながら，収集データは欠損値や外れ値，データ計測機器の測定単位の違いや表記の揺れなどを含んでおり，そのまま分析モデルに入力してしまうと，正しい結果を得られない [1]．例えばセンサーデータの場合，使用するセンサーによる計測間隔や単位の違いに加え，個々の機器の計測誤差や，収集にネットワークを介することによる転送タイムアウトや途中でロストする可能性など，単位の統一や外れ値や欠損値の対処といった「前処理」が必要である．分析システムにおいて，データの前処理には 80% のリソースが費やされているため，分析者の作業量に大きなインパクトを与えている [2]．そのため，分析者の負荷を減らす方法として，前処理の一部を自動化が望まれている．

データ分析を検討する際，データ分析の目的などのシステムの検討からデータの前処理などのデータを準備しモデルを生成，その後のナレッジの共有までの一連のフローを「データマイニング」と呼ぶ．データマイニングに必要な処理を体系立てたフレームワークの一つ APREP-DM（Automated PRE-Processing for Data Mining）では，前処理を自動化可能な処理である “Common processing for Data Cleaning” と，モデルに適した値に加工するため試行錯誤など人手が必要な処理である “Other Processes for Data Cleaning” の 2 つのサブ処理を定義している [3]．本稿では，前処理で行われる作業のうち，自動化可能な Common processing for Data Cleaning に含まれる 外れ値・欠損値の対処をベイズ推論を用いて自動で補完する手法を提案する．さらに，データマイニングフレームワークでは，分析の目的や成功基準などを人の知識により設定する “Business Understanding” が重要とされている．この Business Understanding とベイズ推論を PBE（Programming by Example）アプローチにより融合させ，少ないデー

[1] 東京女子大学 大学院理学研究科

© 2019 Information Processing Society of Japan

タで既存手法より同程度以上の精度，かつツールを使用するより分析者の作業負荷が低い学習モデルを提案する．具体的には，1) 複数の補完手法を予め定義し，2) 補完が必要な場所それぞれに対して定義した補完手法の適正度（確率）を出力，3) 分析者が表示された適正度を確認しながら手法を選択，4) 選択した手法を学習データとしてモデルの再生成，を行う．

本稿の貢献は下記 2 点である．

- PBE アプローチにより，外れ値と欠損値の補完値の算出に人の知識を融合させ，人による作業のメリットを活かしつつ前処理を自動化する．
- 提案手法と既存の補完手法を，提案手法に定義された手法を自動選択した場合と人の知識により選択した場合について比較し，有効性を検証する．

本稿の構成は以下の通りである．まず 2 章で提案手法の関連研究について述べ，3 章で本稿提案手法全体のワークフローと特長やアルゴリズムについて述べる．4 章では，提案手法と既存の手法を比較することによる評価を行い，5 章で本稿のまとめを行う．

2. 関連研究

2.1 欠損値補完手法

欠損値は，MCAR（Missing Completely At Random），MAR（Missing At Random），MNAR（Missing Not At Random）の 3 種類に分類される [4]．MCAR は欠損値が完全にランダムに生じているケースであり，欠損値の発生は観測データにも欠損データに依存しない．すなわち，分析に使いたいデータ項目が全てを揃っている，もしくは欠けている項目があったとしても，そのデータパターンが他のデータから推定できる項目のみであると言える．MAR は，欠損値の発生が完全にランダムではないケースであり，欠損値の発生は観測データに依存する．例えば，社員が TOEIC の点数を自己申告しない場合，その値は欠損値となるが，欠損値の発生に点数の値は依存しない．MNAR は，欠損値の発生が欠損値そのものに依存するケースであり，欠損値の発生は欠損データに依存する．例えば，非喫煙者は喫煙有無を回答したが，喫煙者は回答しなかった場合，欠損値は喫煙有無に依存していることになる．本稿対象データは，ひとつひとつのセンサーから取得した時系列のデータであるため，MCAR とする．欠損値の補完方法として，欠損データ箇所を削除する「リストワイズ法（完全ケース分析とも呼ぶ）」や，取得したデータから代入値を一意に定め代入する「単一代入法」，欠損データの分布を母集団としランダムに抽出したデータを入力値とした異なる代入処理のシミュレーションを複数回行うことによって欠損箇所を置き換える「多重代入法」などがある．リストワイズ法は，データの可視化を行う際，欠損値を削除してしまうと値の間隔がずれてしまう可能性があるため今回の入力方法として適さない．また，多重代入法は，前処理や機械学習モデルの有効性を効率的に確認しながら高精度な予測モデルを構築するパイプライン探索手法 [5] などがあるが，多くのデータサンプルを要するため，通常分析者が保持しているデータは 1 つのデータのため現実的に困難である [4]．単一代入法は，平均値代入法やスプライン補間など関数を定義することにより自動で補完してくれるが，一部のみ代入する値を変更するなどのカスタマイズに対応することはできない．そのため，カスタマイズを可能にする手法が必要である．

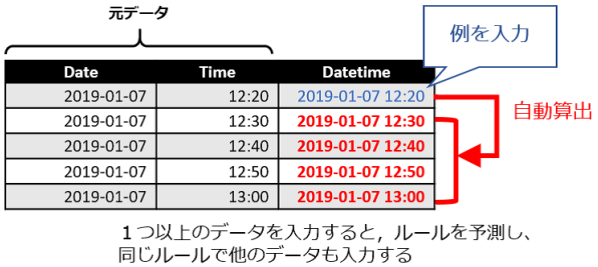

図 1　PBE により，*Date* 列と *Time* 列から *Datetime* 列を生成

2.2 データ変換・加工ツール

分析者の前処理の負荷を減らすための方法として，OpenRefine[6]，Trifacta Wrangler[7] のような前処理の開発をサポートするためのツールが存在する．ツールにより UI は様々であるが，データに外れ値や欠損値があるか否かを確認するためのプロファイリング機能やデータ形式や文字列の統一，データの並び替えや集計などの加工処理を GUI 操作で行うことが可能である．これらのツールは履歴機能を持っており，対象データを変更しても同じ処理であれば，再度処理を開発せずに実行することができる．しかしながら，処理自体は分析者本人がフローを開発する必要があり，さらに処理に変更が生じた場合はメンテナンスを行わなくてはならない．そのため，ツールの使用経験やプログラミングスキルが少ない分析者が使いこなすのは困難である [8]．したがって，分析者がより少ない作業量で正しい処理手順を選択できる手法が求められている．

2.3 PBE

PBE（Programming by Example）アプローチは，1 つ以上の入力と出力の組を元に変換ルールを予測し，自動で加工する手法によるアプローチである．近年，データの値に対する FlashFill[9] や FlashNormalize[10]，データ構造を変換する Foofah[8] など，分析対象データ量の増加により，PBE アプローチはビッグデータのデータ変換手法として注目されている．PBE は 3 つメイン処理：1) 分析者により入力された例と一致するプログラムを効率的に検索することができる検索アルゴリズム，2) 分析者により入力さ

© 2019 Information Processing Society of Japan

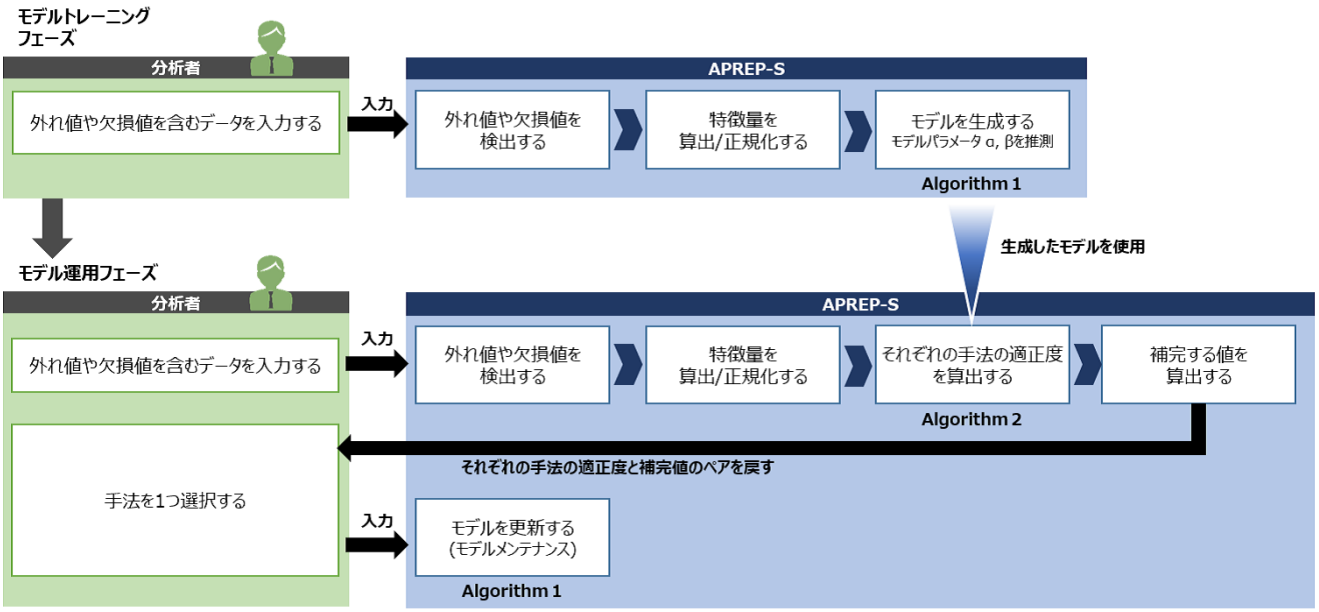

図 2　分析者と APREP-S のワークフロー：緑の箇所は分析者の作業を示し，青は APREP-S の作業を示している．

れた例を満たす処理の中から最適な処理を選択するランク付け，3) ユーザビリティと使いやすさを促進するためのインタラクションモデル，を持つ [2]．PBE の例を図 1 に示す．図 1 は，*Date* 列と *Time* 列を持つ表形式のデータに *Datetime* 列を追加するときの動きを表している．まず，分析者が *Datetime* 列の一番上の行に「2019-01-07 12:20」と入力する．分析者の入力を受け，PBE は入力されたデータのルールが「*Date* 列の後ろに *Time* 列を結合」であることを推測し，2 行目以降に同じルールでデータを入力する．したがって，分析者は 1 番上の 1 行を入力するのみで，それ以下の行の値が補完されたデータを取得することができる．

本稿では，入力と出力の組み合わせ例を学習データとして与えることにより，自動補完では対処できない補完方法のカスタマイズを可能とし，データの外れ値・欠損値を補完する手法の適正度をベイズ推論により提示する手法，APREP-S（Automated PRE-Processing for Sensor data）を提案する．分析者は一例を与えるのみのため，作業量はツールを利用するより少なくなる．

3.　提案手法

分析者が補完したいデータを入力すると，APREP-S があらかじめ設定しておいた複数の手法それぞれに対する適正度と補完値を出力する．分析者は APREP-S が出力した手法の適正度と補完値を確認しながら最適な手法を選択できる．分析者が選択した手法は新たな学習データとしてモデルを更新する．データを分析する際の，分析者と提案手法全体のワークフローを図 2 に示す．提案手法は大きく 2 つのフェーズに分かれる．ひとつがモデルを生成する「モデルトレーニングフェーズ」，そしてもう一つが生成したモデルを実際に使い分析者に適正度と補完値を返す「モデル運用フェーズ」である．

以下，モデルトレーニングフェーズとモデル運用フェーズについて詳述する．

3.1　提案モデル：APREP-S

3.1.1　モデルトレーニングフェーズ

モデルトレーニングフェーズでは，最初に，分析者は外れ値や欠損値を含む学習データを APREP-S に入力する．APREP-S は分析者からデータを受け取った後，外れ値や欠損値の場所を特定し，データ補完ターゲットの特徴 $x_1, x_2, ... \in \boldsymbol{x}$（$\boldsymbol{x}$ の数は有限）を求める．$\boldsymbol{x}$ は正規化された値である．その後，正規化した特徴量を元にモデルを生成する．ベイズ推論を行う対象は，APREP-S に定義されている手法 $m_1, m_2, ... \in \boldsymbol{M}$($m$ の数は有限) である．m は離散値のため，各手法の事後確率 $p(m_k|\boldsymbol{y})$ は，

$$p(m_k|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|m_k)p(m_k)}{\sum_{i=1}^{K} p(\boldsymbol{y}|m_i)p(m_i)} = \frac{\exp(\boldsymbol{y}(x_k))}{\sum_{i=1}^{K} \exp(\boldsymbol{y}(x_i))} \quad (1)$$

$$\boldsymbol{y}(x_q) = \alpha + \boldsymbol{\beta}\, x_q \quad (1 \leq q \leq Q) \quad (2)$$

となる．K は APREP-S で定義されている手法の数，Q は APREP-S で定義されている特徴の数を示している．APREP-S のモデルは 2 つのパラメータ α と $\boldsymbol{\beta}$ を持つソフトマックス関数 $\boldsymbol{y}(x)$ とした．さらに，尤度関数はカテ

© 2019 Information Processing Society of Japan

Algorithm 1 APREP-S モデルの生成（モデルトレーニングフェーズ）

INPUT: 手法番号リスト（tr_m）
正規化済みの特徴量（$\boldsymbol{x}$）
OUTPUT: APREP-S model

$\alpha \leftarrow \mathcal{N}(\mu_\alpha, \sigma_\alpha)$
$\boldsymbol{\beta}$ （Q-by-K 行列（*））$\leftarrow \mathcal{N}(\mu_\beta, \sigma_\beta)$
for $k \leftarrow 1$ *to* K **do**
  for $q \leftarrow 1$ *to* Q **do**
    $\boldsymbol{y} = \alpha + \boldsymbol{\beta}\, x_q \leftarrow \alpha, \boldsymbol{\beta}$
    $p(m_k|\boldsymbol{y}) = \exp(\boldsymbol{y}(x_q)) / \sum_{i=1}^{K} \exp(\boldsymbol{y}(x_i)) \leftarrow m_k, \boldsymbol{y}$
    $p(\boldsymbol{y}|m_k) \leftarrow tr_m$
    $C(m_k|\boldsymbol{y}) \leftarrow \boldsymbol{y}, p(\boldsymbol{y}|m_k)$
    $\alpha_p, \boldsymbol{\beta}_p \leftarrow$ sampling with $C(m_k|\boldsymbol{y})$
  end for
end for
$\boldsymbol{y} \leftarrow \alpha_p, \boldsymbol{\beta}_p$
APREP-S model $\leftarrow \boldsymbol{y}$

(*) K : APREP-S に定義されている手法の数
Q : APREP-S に定義されている特徴量の数

ゴリカル分布で与えられるため，

$$C(\boldsymbol{M}|\boldsymbol{y}) = \prod_{k=1}^{K} (y_k^{u_k}) \tag{3}$$

となる．ここで，u_k は m_k ($0 \leq y_k \leq 1$ かつ $\sum_k y_k = 1$) の確率分布である．$p(m_k|\boldsymbol{y}) = \sum_{i=1}^{K} u_i = 1$ であるため，$p(m_k|\boldsymbol{y})$ は正規化された指数関数である．したがって，各手法の適正度は $p(m_k|\boldsymbol{y})$ と同じになる．APREP-S のモデル生成処理のアルゴリズムは Algorithm 1 に示す．APREP-S モデルのパラメータ α は，APREP-S で定義している手法数の要素を持つ配列であり，パラメータ $\boldsymbol{\beta}$ は，APREP-S で定義している特徴数の行と APREP-S で定義している手法数の列を持つ行列である．α と $\boldsymbol{\beta}$ の確率分布はガウス分布とする．Eq.（1），Eq.（2），Eq.（3）を使い，APREP-S モデル $\boldsymbol{y} = \alpha_p + \boldsymbol{\beta}_p \boldsymbol{x}$ を求める．α_p と $\boldsymbol{\beta}_p$ は α と $\boldsymbol{\beta}$ の事後分布である．

モデルトレーニングフェーズはモデルを生成するまでであり，生成したモデルは APREP-S 内で保持するため，分析者に戻す値はない．

3.1.2 モデル運用フェーズ

モデル運用フェーズでは，最初に，分析者はデータ補完を行いたい外れ値や欠損値を含むデータを APREP-S に入力する．APREP-S は分析者からデータを受け取った後，外れ値や欠損値の場所を特定し，データ補完ターゲットの特徴 $x_1, x_2, ... \in \boldsymbol{x}$ を求める．$\boldsymbol{x}$ はモデルトレーニングフェーズの特徴と同じであり，正規化された値である．正規化した特徴量を使用し，APREP-S で定義された手法 $m_1, m_2, ... \in \boldsymbol{M}$ それぞれに対し，適正度を算出する．この時，モデルトレーニングフェーズで生成したモデルを使用する．適正度を算出するアルゴリズムは Algorithm 2 に示す．モデルのパラメータ α と $\boldsymbol{\beta}$ は確率分布がガウス分布のパラメータのため，期待値を固定値とする．APREP-S モデルに固定化したパラメータを代入し，入力したデータの補完場所に対する適正度を計算した後，最後に百分率に変換する．適正度を求めた後は，APREP-S で定義している手法それぞれについて，補完する値を計算する．その後，適正度と補完値のペアを分析者に返す．分析者は，APREP-S から受け取った適正度と補完値を確認し，それぞれの補完場所に対し，補完する値を選択する．選択した手法は新たな学習データとし，APREP-S モデルを更新する．

Algorithm 2 各手法の適正度算出 (モデル運用フェーズ)

INPUT: 正規化済みの特徴量（$\boldsymbol{x}$）
Algorithm 1 で生成した APREP-S モデル
APREP-S のパラメータ（$\boldsymbol{\alpha}_p, \boldsymbol{\beta}_p$）
OUTPUT: 各手法の適正度(単位:%)

$\boldsymbol{y} \leftarrow$ APREP-S モデル
for $k \leftarrow 1$ *to* K **do**
  for $q \leftarrow 1$ *to* Q **do**
    $\mathbb{E}_\alpha[\alpha|m_k] = \sum_q p(\alpha_p|m_k)\alpha_p$
    $\alpha \leftarrow \mathbb{E}_\alpha[\alpha|m_k]$
    $\mathbb{E}_\beta[\boldsymbol{\beta}|x_q, m_k] = \sum_q p(\boldsymbol{\beta_p}|m_k, x_q)\boldsymbol{\beta_p}$
    $\beta \leftarrow \mathbb{E}_\beta[\boldsymbol{\beta}|x_q, m_k]$
  end for
  $\boldsymbol{y} = \alpha + \beta\boldsymbol{x} \leftarrow \alpha, \beta$
  $p(m_k|\boldsymbol{y}) = \exp(\alpha + \beta\boldsymbol{x}) / \sum_{k=1}^{K} \exp(\alpha + \beta\boldsymbol{x}) \leftarrow \boldsymbol{x}$
  $P_k(m_k|\boldsymbol{y}) = 100 * p(m_k|\boldsymbol{y})$
end for

(*) K : APREP-S に定義されている手法の数
Q : APREP-S に定義されている特徴量の数

4. 評価

本評価では，モデルトレーニングフェーズで 2 種類のモデルを生成する．ひとつが機械学習を想定した「手法を自動選択した学習データにより生成したモデル」，もうひとつが分析者が手法を選択することを想定した「人の知識により選択したデータにより生成したモデル」である．2 つのモデルを生成した後，生成したそれぞれのモデルに対し，モデル運用フェーズで出力した補完値と，既存の補完方法で求めた補完値を比較する．使用するデータは，オリジナルデータに外れ値と欠損値を入れ込んだデータを生成する．各手法の精度は，オリジナルデータと APREP-S または既存の補完方法で求めた値との二乗和誤差

$$ERR = \frac{1}{2} \sum_{i=1}^{I} (org_i - v_i)^2 \tag{4}$$

を比較する．ここで，org_i は i 番目のオリジナルデータ，v_i は i 番目の APREP-S，または既存手法で算出した補完値である．ERR が小さいほどオリジナルデータに近い値を補完できていることになるため，精度が高いと判断する．

4.1 データセット

ワイヤレスの気温と湿度センサー（DHT-22）のデータを

© 2019 Information Processing Society of Japan

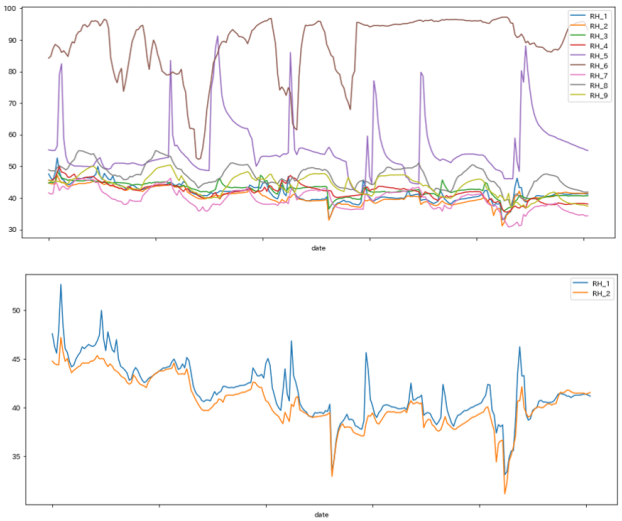

図 3 湿度（RH）データの 1 週間の時系列グラフ：上の図は全ての RH データのグラフ，下の図は $RH1$ と $RH2$ のグラフ（青が $RH1$，オレンジが $RH2$ を示す）

含むデータセットをオリジナルデータとして使用する [11]．データセットは 29 列あり，計測時間，気温，湿度，気圧，風速などのデータが格納されている．気温と湿度は，キッチン，リビング，寝室，バスルームなど，室内外 9 箇所に設置されたセンサーにより収集している．データ件数は，137 日（4.5 ヶ月）のデータであり，各センサーごとに 19,735 件である．センサーからは 3.3 分ごとにデータが収集されるが，本データセットは 10 分ごとに集計したデータセットとなっている．また，センサー DHT-22 の気温の誤差は ± 0.5°C，湿度の誤差は ± 3% である．

本評価では，キッチンエリアの湿度データ $RH1$ とリビングエリアの湿度データ $RH2$ を選択し，モデルトレーニングフェーズの入力データを $RH2$，モデル運用フェーズの入力データを $RH1$ とする．RH データのグラフを図 3 に示す．上の図は当該データセットに格納されている全ての湿度 RH データのグラフである．$RH5$ はバスルームのデータ，$RH6$ は屋外のデータであり，他のデータと明らかにグラフの形状が異なることがわかる．本評価では，$RH5$，$RH6$ 以外のデータから波形が似ている $RH1$ と $RH2$ を選択した．図 3 の下の図は $RH1$ と $RH2$ のグラフであり，青が $RH1$，オレンジが $RH2$ を示している．

評価用として，選択した $RH1$ と $RH2$ に外れ値と欠損値を挿入する．外れ値もしくは欠損値が起こる確率は指数分布 Eq.（5）に従うものとし，9 回に 1 回欠損値が生じると仮定した．外れ値の外れ度合いはガウス分布 Eq.（6）に従うとする．

$$f(e) = \frac{1}{\epsilon}\exp\left(-\frac{e}{\epsilon}\right) \quad (500 \leq \epsilon \leq 1000) \tag{5}$$

$$f(e) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(e - org_i)^2}{2\sigma^2}\right\} \quad (0 \leq \sigma^2 \leq 10) \tag{6}$$

ここで，σ^2 はガウス分布の分散を示している．結果として，$RH1$ は 36 箇所，$RH2$ は 38 箇所の補完箇所を含むデータを生成した．

4.2 APREP-S 内の補完値算出手法

本評価では，3 つの補完手法 $m_1, m_2, m_3 \in \boldsymbol{M}$ を APREP-S に定義した．m_1 は補完箇所前後の平均値を代入，m_2 は直前の値をそのまま代入，m_3 はスプライン補間による代入とした．o_i を i 番目のデータとすると，m_1, m_2, m_3 は以下となる．

- $m_1(o_i) = (o_{i-1} + o_{i+1})/2$
- $m_2(o_i) = o_{i-1}$
- $m_3(o_i) = a_j(o_i - o_j)^3 + b_j(o_i - o_j)^2 + c_j(o_i - o_j) + d_j$ $(1 \leq j \leq N-1)$

4.3 特徴量

本評価では，3 つの特徴量 $x_1, x_2, x_3 \in \boldsymbol{x}$ を定めた．x_1 は補完箇所前後の傾き，x_2 は補完箇所の直前 2 つのデータの傾き，x_3 は補完箇所前後の平均値とデータ全体の平均値との差とする．o_i を i 番目のデータとすると，x_1, x_2, x_3 は以下となる．

- $x_1(o_i) = (o_{i-1} + o_{i+1})$
- $x_2(o_i) = (o_{i-2} + o_{i-1})$
- $x_3(o_i) = |(o_{i-1} + o_{i+1})/2| - \mathrm{mean}(\boldsymbol{o})$

4.4 モデル評価

4.4.1 モデルトレーニングフェーズ

分析者が $RH2$ を APREP-S に入力すると，APREP-S はまず補完箇所，すなわち外れ値または欠損値を検出する．その後，それぞれの補完箇所に対する正規化した特徴量 x_1, x_2, x_3 を計算する．この特徴量を元に，APREP-S モデルを生成する．本評価では，APREP-S モデルのパラメータ α と $\boldsymbol{\beta}$ はガウス分布（平均:μ=0，分散:σ^2=2) に従うとした．また，初期モデルを生成するにあたり，本評価では 2 つの学習データ，（1）特徴量から選択する手法を自動で選択する方法，（2）人の知識により手法を選択する方法，を基に 2 つのモデルを生成する．

まず，（1）特徴量から選択する手法を自動で選択する方法では，$x_2 > 0.05$ すなわち直前の 2 つのデータの差がある場合は m_1 を，$x_2 < 0.05$ かつ $x_3 < 0.05$ すなわち補完箇所直前の 2 つのデータに差がなくかつデータの平均に近い場合は m_2 を，それ以外は m_3 とした．入力データ $RH2$ の 38 の補完箇所に対して選択した手法は時系列の順番に $(1, 3, 3, 3, 1, ..., 2, 3, 1, 1)$（38 箇所）となり，これを学習データとする．

次に，（2）分析者が手法を選択する方法では，モデル更新時と同様，補完箇所に各手法で補完する代入値を計算したうえで，入力データ $RH2$ の 38 の補完箇所に入力する手

© 2019 Information Processing Society of Japan

法番号を選択した．（1）の自動で選択する方法と比較し，15 箇所が異なる学習データとなった．

特徴量 $\boldsymbol{x}$ と手法 $\boldsymbol{M}$，そして（1）（2）それぞれの選択した手法番号リストを使い，Algorithm 1 により，APREP-S モデルを生成する．

4.4.2 モデル運用フェーズ

モデルトレーニングフェーズで生成した APREP-S モデルを使い，手法の適正度を推測する．生成した 2 つのモデルそれぞれに対し，$RH1$ の補完を推測すると，以下の結果が得られる．ただし，以下は例としてデータは最初の 3 箇所のみであり，括弧内の値は補完値を示している．まず，（1）特徴量から選択する手法を自動で選択する方法は，

1st: $m_1 = 62\%(44.7)$, $m_2 = 1\%(42.6)$, $m_3 = 37\%(42.6)$
2nd: $m_1 = 91\%(44.2)$, $m_2 = 1\%(40.8)$, $m_3 = 8\%(40.8)$
3rd: $m_1 = 3\%(42.7)$, $m_2 = 2\%(41.6)$, $m_3 = 95\%(41.6)$

となる．次に，（2）分析者が手法を選択する方法は，

1st: $m_1 = 96\%(44.7)$, $m_2 = 3\%(42.6)$, $m_3 = 1\%(42.6)$
2nd: $m_1 = 98\%(44.2)$, $m_2 = 1\%(40.8)$, $m_3 = 1\%(40.8)$
3rd: $m_1 = 32\%(42.7)$, $m_2 = 59\%(41.6)$, $m_3 = 9\%(41.6)$

となる．本評価では，一番高い適正度の手法を採用する．例えば，（1）の場合は最初の 3 箇所の手法は手法 1→ 手法 1→ 手法 3 と選択し，（2）は手法 1→ 手法 1→ 手法 2 と選択することとする．

4.5 比較する既存手法

APREP-S と比較する手法は, (i) 全ての補完箇所にデータ全体の平均値を代入 $= \mathrm{mean}(\boldsymbol{o})$, (ii) 補完箇所前後 36 個ずつのデータの平均値を代入 $= \mathrm{mean}(o_j)$ $(i-36 \leq j \leq i+36)$, (iii) 3 次スプライン補間$=a_j(o_i-o_j)^3+b_j(o_i-o_j)^2+c_j(o_i-o_j)+d_j \quad (i-36 \leq j \leq i+36)$ の 3 つとした. ここで，$\boldsymbol{o}$ は $RH1$ データ，o_i は $RH1$ の i 番目のデータとする．今回，評価に使用しているデータセットは 10 分間隔のデータのため，36 個は 6 時間分のデータである．したがって，(ii) と (iii) は，前後 6 時間，合計 12 時間分のデータを対象としている．

4.6 結果

評価結果を表 1 に示す．自動選択の場合でも，人による選択の場合でも，APREP-S が一番精度が高い結果となった．自動選択と人による選択の比較では，人による選択の方が精度が 3 倍であった．既存の補完手法と比較し，本稿が対象としているセンサーデータの補完において，APREP-S の有効性が検証できた．

5. おわりに

本稿では，センサーデータのように，外れ値や欠損値を含むデータの補完・変換を行うために，PBE アプローチにより人の知識と機械学習を融合した APREP-S を提案した．本稿の結論は以下の 2 点である．

- 機械学習を用いた自動処理により分析者の作業負荷削減と精度を，人の知識を機械学習に入力することによりカスタマイズ性を，併せ持つ手法を提案した．分析者のカスタマイズ性を高めるため，APREP-S に定めた手法の適正度と補完値を出力し，PBE アプローチによりモデルに入力する手法を定めた．
- 自動生成による学習データと人の知識を入れた学習データの 2 種類のデータに対し，既存の補完値手法と提案手法 APREP-S で補完した結果の精度を比較し，APREP-S の有効性を検証した．

本稿の検証では，目視で学習データと運用データを選択した．今回の検証では APREP-S が有効であるという結果が出たが，学習データが適切であったか否かの検証が不十分である．また，実運用を想定すると，学習データの選択は分析者が選択することになる．そのため，適切な学習データを自動で選択する手法を組み込むことができれば，当該処理の作業負荷を削減するだけではなく，より精度の高い結果を得ることができると考える．さらに本稿では，外れ値や欠損値は連続していないデータを用いている．しかしながら，ベイズ推論や機械学習を利用すれば，一定範囲のデータが欠損している場合やデータの粒度を 1 時間ごとを 10 分毎に変換するなどデータの粒度を小さくする場合など，補完箇所が連続している場合にも応用できると考えている．次のステップとして，連続した補完箇所の補完手法を研究・検証予定である．

表 1 二乗和誤差（Eq.（4））による評価結果

手法選択方法	APREP-S	Single	Multi	Spline
自動選択	355.49	935.75	683.26	526.80
人による選択	117.10	935.75	683.26	526.80

(*) “Single”：全ての補完箇所にデータ全体の平均値を代入
“Multi”：補完箇所前後 36 個ずつの平均値を代入
“Spline”：3 次スプライン補間による代入

参考文献

[1] Qi, Zhixin and Wang, Hongzhi and Li, Jianzhong and Gao, Hong: Impacts of Dirty Data: and Experimental Evaluation, *arXiv:1803.06071 [cs, stat]* (2018).

[2] Gulwani, Sumit and Jain, Prateek: Programming by Examples: PL meets ML, *Microsoft Research*, (online), available from ⟨https://www.microsoft.com/en-us/research/publication/programming-examples-pl-meets-ml/⟩ (2017).

[3] Hiroko Nagashima and Yuka Kato: APREP-DM: a Framework for Automating the Pre-Processing of a Sensor Data Analysis based on CRISP-DM, *PerFoT'19 - International Workshop on Pervasive Flow of Things (PerFoT'19)*, Kyoto, Japan (2019).

[4] Graham, John W.: Missing Data Analysis: Making It Work in the Real World, *Annual Review of Psychology*, Vol. 60, No. 1, pp. 549–576 (online), DOI: 10.1146/annurev.psych.58.110405.085530 (2009).

© 2019 Information Processing Society of Japan

[5] 塩田哲哉, 及川一樹, 澤田雅人：サンプリングを用いた機械学習パイプライン探索手法，一般社団法人人工知能学会，pp. 3Q17in1–3Q17in1 (2017).

[6] Qi Jacky Cui, Thad Guidry, Martin Magdinier: OpenRefine, Supported by Google News Initiative (online), available from ⟨http://openrefine.org/⟩ (accessed 2019-05-05).

[7] Trifacta: Trifacta Wrangler, Trifacta (online), available from ⟨https://www.trifacta.com/start-wrangling/⟩ (accessed 2019-05-05).

[8] Jin, Zhongjun and Anderson, Michael R. and Cafarella, Michael and Jagadish, H. V.: Foofah: Transforming Data By Example, *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, ACM, pp. 683–698 (online), DOI: 10.1145/3035918.3064034 (2017).

[9] Gulwani, Sumit: Automating String Processing in Spreadsheets Using Input-output Examples, *Proceedings of the 38th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '11, ACM, pp. 317–330 (online), DOI: 10.1145/1926385.1926423 (2011).

[10] Kini, Dileep and Gulwani, Sumit: FlashNormalize: Programming by Examples for Text Normalization, *IJCAI*, pp. 776–783 (online), available from ⟨https://www.microsoft.com/en-us/research/publication/flashnormalize-programming-examples-text-normalization/⟩ (2015).

[11] Candanedo, Luis M. and Feldheim, Veronique and Deramaix, Dominique: Data driven prediction models of energy use of appliances in a low-energy house, *Energy and Buildings*, Vol. 140, pp. 81–97 (online), DOI: 10.1016/j.enbuild.2017.01.083 (2017).

© 2019 Information Processing Society of Japan