
一般口演

一般口演26

機械学習・HISデータ解析

2017年11月23日(木) 09:15 ～ 10:45 F会場 (10F 会議室1004-1005)

[4-F-1-OP26-4] 状態空間モデルと深層ニューラルネットワークによる検体検査結果の欠損値推定精度の比較河添 悦昌¹, 倉沢 央², 岩井 聡¹, 香川 璃奈¹, 大江 和彦¹ (1.東京大学大学院 医学系研究科 医療情報学分野, 2.日本電信電話株式会社)

【背景・目的】医療における時系列データの特徴として、測定間隔が不規則であること、同時に全項目が測定されないことが挙げられ、そのため生じる欠損値が時系列解析の上で問題となる。本研究は検体検査結果の欠損を精度よく推定することを目的とし、状態空間モデル（SSM）と2種類のニューラルネットワーク（NN）を用いた方法を比較する。

【方法】入院患者23,920人を対象とし、入院初日から28日目までの検体検査3項目（WBC、Hb、PLT）の結果を抽出し、23,920人分の3項目×28日のデータを得た。以下の方法による欠損値の推定精度を比較した。1) 状態空間モデル。SSMは時系列を観測値と状態値に分けてモデル化し、観測値が欠損する場合にも同時点の状態の値から観測値を推定する。2) AutoEncoder（AE）。AEは入力と出力を同値にするよう学習するNNであり、入力にノイズを与えた場合もそれを復元することに着想を得て、時系列の欠損値推定に応用する。3) Convolution-Deconvolution（CD）。CDは1次元の入力に対して畳込みを繰り返し、同回数の逆畳込み込みから入力を復元するように学習するNNである。評価においては、モデル構築後に非欠損値を1件ずつマスクして推論を行い、推論値と真値との誤差の絶対値を算出し比較した。

【結果】以下Hb系列の絶対誤差（単位はg/dL）を示す。平均値、中央値、最大値はそれぞれ、SSMが0.58、0.42、14.65、AEが0.67、0.52、14.35、CDが0.54、0.40、14.51であった。他の2つの系列においても、CDによる方法が他に比べ低い誤差を示す傾向にあった。

【考察】いずれの方法も平均値と中央値に乖離があり、いくつかの大きくはずれた値が平均値を上昇させていた。このことは、実際の利用に影響を及ぼす程度と思われ精度の改善が必要である。

状態空間モデルと深層ニューラルネットワークによる 検体検査結果の欠損値推定精度の比較

河添 悦昌^{*1}、倉沢 央^{*2}、岩井 聡^{*1}、香川 璃奈^{*1}、大江 和彦^{*1}

^{*1} 東京大学大学院医学系研究科 医療情報学分野 ^{*2} 日本電信電話株式会社

Comparison of Missing Value Estimation of Laboratory Tests between State Space Model and Deep Neural Network

Yoshimasa Kawazoe^{*1}, Hisashi Kurasawa^{*2}, Satoshi Iwai^{*1}, Rina Kagawa^{*1}, Kazuhiko Ohe^{*1}

^{*1} Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo

^{*2} NTT

[Background] Since time-series data in medical care including irregular interval and inability of simultaneous measurement of all items cause missing value, it often becomes a problem for time series analysis. To achieve better estimation of missing value in laboratory test results, methods between those based on state space model and two types of neural networks (NN) are compared. [Methods] Three kinds of laboratory test results (WBC, Hb, and PLT) for 25,272 inpatients conducted from their hospitalization to day 28 was obtained. Estimation of missing value was compared between the following methods. 1) State space model (SSM): Since it is modeled by dividing into observation and state in time series, observation is estimated based on a state at the same point in time even if the observation was missing. 2) Auto Encoder (AE): AE is an NN which learns to make input and output to be the same. It is applied to missing value estimation in time series based on a concept to restore input value even if a noise was given. 3) Convolution-Deconvolution (CD): CD is an NN which learns to restore input by repeating convolution by multiple times for one-dimensional input and repeating deconvolution by the same times. In the evaluation, an absolute error between an estimate and a true value was calculated by masking each non-missing value after building up a model. [Results] Absolute errors of Hb (unit:g/dL) are shown below. Average, mean, and maximum values of errors were 0.58, 0.42, and 14.65 in SSM, 0.56, 0.42, and 8.32 in AE and 0.54, 0.40, and 8.48 in CD, respectively. The method by CD tended to show lower error compared with those of other methods also in other two systems. [Discussions] With a gap between average and mean values observed in all methods, several greatly outliers pushed up the average values. Since it might affect actual use of those methods, further improvement is required.

Keywords: EHR, Laboratory Test, Missing Value Estimation, State Space Model, Neural Network.

1. 緒論

電子カルテデータは longitudinal なデータを多く含み、これは 1 つの対象について経過を追って複数の時点で観察結果を記録したものである。特に血液や尿などの検体検査結果は典型的な longitudinal データであり、多系列であること、測定間隔が不規則であること、同時に全項目が測定されないなどの理由により、時系列データとして扱う際に欠損値を生じ、解析の上での問題となる。そのため、データマイニングや機械学習による利用の前処理として何らかの方法で欠損値を埋める必要が生じる。検体検査結果は測定間隔が不等であることから、その欠損値を補間する際には間隔が一定であることを前提としない線形補間やガウス過程によるモデル化が用いられるが、これらの方法は単系列を対象とするものである。しかしながら、検体検査結果は複数の検査項目、つまり多系列で構成されることから、ターゲットとなる系列以外の系列の情報を使うことで、欠損値の推定精度がより向上すると期待される。

2. 目的

本研究は一般病棟に入院する患者の検体検査結果の欠損、つまり日の単位で生じる欠損を精度よく推定することを目的とし、状態空間モデル(SSM)と 2 種類のニューラルネットワーク(NN)を用いた方法において、単系列の情報のみを使用するモデル(以下、単系列モデルと呼ぶ)と、多系列の情報を使用するモデル(以下、多系列モデルと呼ぶ)の欠損値の推定精度の比較、ならびに欠損値の数による精度の比較を行う。

3. 方法

3.1 対象データ

東京大学医学部附属病院の入退院記録から 28 日以上入院期間を有する患者を抽出した。各患者において入院初日から 28 日目までの 3 項目の検体検査(白血球数:WBC、ヘモグロビン値:Hb、血小板数:PLT)の結果を抽出した。28 日の期間中に検査が行われた日数が 4 日以下の患者を除外し、25,272 人の 3 項目×28 日間のデータを得てこれを対象とした。対象データセットがどの程度欠損を有するかを示すために、各検査項目について、28 日の期間に検査が行われた日数と症例頻度との関係を図 1 に示す。各データは最低で 0 日、最大で 23 日の欠損を有し、全体の半数以上が 14 日以上欠損を有するものであった。

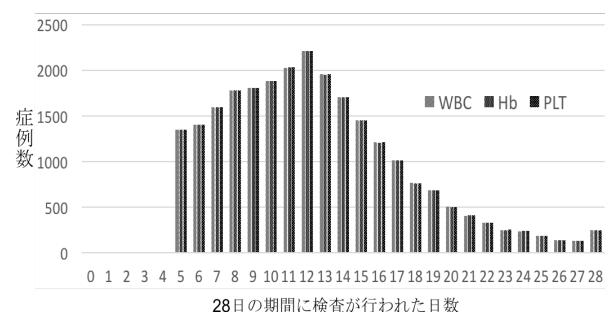


図 1 28 日間に検査が行われた日数と症例数

WBC、Hb、PLT の総検査回数はそれぞれ、314,963、314,940、314,921 であり、同日の検査ではほぼ 3 項目が同時に測定されていると考えられた。同日に複数回の検査が行われているものは平均値をその日の値とした。対象データの要約統計を表 1 に示す。解析の際には各系列の平均値はそのままとし、分散を 1 に標準化したものを用いた。

表1 対象データの要約統計

項目名	平均	中央	標準 偏差	総回数	欠損 割合
WBC (10 ³ /uL)	7.84	6.8	6.52	314,963	55%
Hb (g/dL)	10.6	10.4	2.17	314,940	55%
PLT (10 ⁴ /uL)	22.8	21.2	13.4	314,921	55%

3.2 状態空間モデルを用いた欠損値推定

SSM は時系列データを観測値と状態値に分けてモデル化するため、観測値が欠損する場合にも同時点の状態の値から観測値を推定することができる。¹⁾本研究は、SSM の基本形であるローカルレベルモデルを用いて欠損値の推定を行った。単変量時系列の時点 t の観測値を y_t とし、同じく時点 t の状態値を α_t とする。状態 α_t から観測値 y_t を得る際には観測ノイズ ε_t が加わるとする。また、状態の値は時間に対してランダムウォークにより遷移、つまり α_t は 1 つ前の α_{t-1} からシステムノイズ η_t だけ動くとする。観測ノイズ ε_t とシステムノイズ η_t はそれぞれ、平均 0、分散 σ_ε^2 、 σ_η^2 に従うと仮定し、観測値 y が与えられた際の尤もらしい状態 α の分布を推定することでモデルを構築する。多系列モデルへの拡張は、各系列のシステムノイズ、観測ノイズの間に相関をもたせることで行う。

SSM の推定問題は 3 つに大別され、1 つは途中時点 t までの観測値を用いて当該時点の α_t を推定するフィルタリング、2 つ目は与えられた観測値を用いて各時点の状態 $\alpha_1, \alpha_2 \dots \alpha_t$ を推定する平滑化、3 つ目はフィルタリングの最終時点を先に進め将来の状態と観測値についての予測である。ここで、欠損値の補間は平滑化の問題として扱うことができ、状態 $\alpha_1, \alpha_2 \dots \alpha_n$ と観測値 $y_1, y_2 \dots y_n$ の同時分布を考え、 y の一部が与えられたもとで各状態の条件付き分布を求める問題として扱えば、多変量正規分布に基づく行列計算で推定できる。実際には、時点数 n が大きくなるにつれて計算量が急激に増加し時間的に計算不能となるため、本研究はカルマンフィルタと呼ばれる状態の条件付き分布を逐次的に計算するアルゴリズムを用いて行った。²⁾

3.3 Autoencoder を用いた欠損値推定

Autoencoder(AE)の基本形は3層のNNにおいて、入力層と出力層を同じ値とするようにネットワークを学習するものである。³⁾AE は主に入力データの次元削減や異常データの検知⁴⁾に用いられるが、本研究では入力データにノイズを与えた場合でもそれを復元する Denoising Autoencoder⁵⁾に着想を得て、時系列データの欠損値推定に応用する。ネットワークの学習時は入力データの何割かをランダムに選択し(以下、この割合を Corruption Rate:CRと呼ぶ)、選択した値に0を代入した上で順方向に推論し、0を代入した値を含む非欠損値に対応する推論値と真値との平均絶対誤差を最小化するようにネットワークを更新する。ここで、真の欠損値に対応する推論値は評価出来ないため誤差の計算に含めない。単系列のモデルにおいて、1入力データは1患者における固定日数分の時系列データによって構成することとし、今回の実験では 1 次元 28 要素のデータを入力とした。多系列のモデルは、各系列を直列に連結したものを入力とし、ターゲットとする系列のみを出

力することとし、今回の実験では 1 次元 84 要素を入力とし、出力データの要素数を 28 とした。図 2 に AE により欠損値を推定するネットワークの概要を示す。説明のため中間層を 1 つとしているが、実際の実験ではネットワークを多層にした。

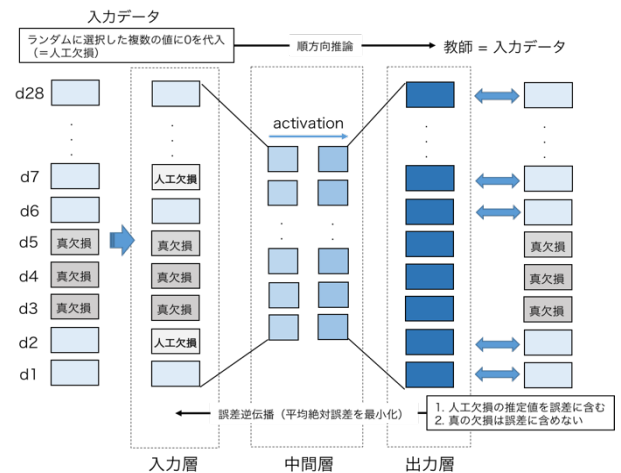


図2 AEにより欠損値を推定するネットワークの概要図
単系列モデルにおける概略を示す。多系列モデルは各系列データを直列につなぎ要素数 84 のデータを入力とした。

3.4 Convolution-Deconvolution を用いた欠損値推定

ニューラルネットワークによる逆畳み込み処理⁶⁾は、不鮮明な画像の復元や物体の輪郭検出などに応用される技術である。本研究では畳み込みと逆畳み込みを用いるニューラルネットワークによる欠損値推定モデルを Convolution-Deconvolution (CD)と呼ぶことにする。CDは1次元の入力に対してフィルタによる畳み込みと逆畳み込み処理を行い、入力を復元するように学習する。AEと同様の考え方から、入力データにノイズを加えてもそれを復元するようにCDを学習することで、時系列データの欠損値推定に応用する。AEと同様に、ネットワークの学習時は元データのいくつかをランダムに選択し、この値に0を代入したものを入力データとする。入力データは1次元のフィルタによる畳み込みによって、より小さい要素数のデータに圧縮し、逆畳み込み層ではこの圧縮されたデータ間を埋めた後にフィルタによる畳み込みを行い、入力と同じ要素数のデータを復元する。0を代入した値を含む非欠損値に対応する推論値と真値との平均絶対誤差を最小化するように畳み込みフィルタを更新することで学習を進める。画像をはじめとする2次元のデータに対しては、畳み込み・逆畳み込みの前後にプーリング処理を設ける場合が多いが、本研究の対象とする系列データは位置特性の吸収がそれほど有効ではないと思われるため、プーリング処理を省略した。単系列のモデルは1次元28要素のデータを1チャンネルで入力し同要素数のデータを出力した。多系列のモデルは各系列1次元28要素のデータを3チャンネルで入力し、ターゲットとする系列の1チャンネル28要素のデータを出力した。図3にCDによって欠損値を推定するネットワークの概要図を示す。実際の実験ではAEと同様にネットワークを多層にしたものを用いた。

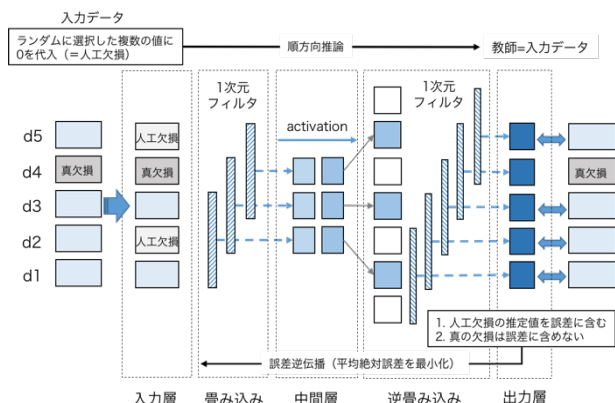


図3 CDにより欠損値を推定するネットワークの概要図

説明のため入力データの要素数を5としているが実際の要素数は28である。また、多系列のモデルは各系列を重ね3チャンネルとして入力する。

3.5 実験設定

3.5.1 SSMのモデル構築と評価方法

WBC、Hb、PLTの各系列における各データの非欠損値をランダムに1つ選択し、これをマスク(欠損値を代入)したものを入力としてSSMを構築し、モデルの推定値と真値との絶対誤差を評価値とした。つまり、真の欠損値に加えて人工的な欠損を1つ発生させ、この欠損に対する推定値の誤差を評価したことになる。また、1データに対してマスクする数(以下、この数をNumber of Missing:NMと呼ぶ)を2と3とした場合の評価も行った。実装はR3.3.2で行い、SSMのライブラリはKFASを用いた。

3.5.2 NNモデルの構築と評価方法

AEとCDのモデル構築と評価は2分割交差検定(学習用50%、テスト用50%)で行った。モデルの構築は学習用データセットを使い、前述の方法により検査系列(WBC、Hb、PLT)毎に単系列と多系列のモデルをAEとCDで各6個、計12個構築した。評価時は、テスト用データセットにおける各データから非欠損値をランダムに1つ選択し(NM=1)、この値に0を代入した上で学習済みのモデルで推論し、推論値と真値との平均絶対誤差を評価した。SSMと同様に、NM=2、NM=3の場合の評価も行った。

AEとCDのハイパーパラメタは経験的に決定した。ネットワークの構造に関して、AEは多層構造とし中間層は5層とした。中間層のユニット数は、単系列のモデルは各28、多系列のモデルは各84とした。CDも多層構造とし中間層は7層とした。フィルタの枚数とサイズはそれぞれ64枚~192枚、7~2とし、最初の入力と最後の出力の両端に1のPaddingを行った。学習時のハイパーパラメタについて、活性化関数はReLU、オプティマイザはAdam、ミニバッチサイズを100とし、1500エポック(学習データセットを1500回繰り返し入力)終了の時点でモデルを保存し、このモデルを使い評価を行った。実装はPython3系で行い、NNのライブラリはChainer 1.21.0を用いた。

3.5.3 実験の手順と結果の示し方

AEとCDに関してはモデル構築時のCRを決める必要がある。そのため、以下の結果4.1ではCRを0%、25%、50%、75%とした場合のAEとCDの学習曲線を示し、結果4.2では各CRにおけるモデル構築時の誤差と、NM=1とした場合の評価

値を示しCRを決める。結果4.3では本研究の主目的である各手法における欠損値推定の誤差を示す。ここでは、SSM、AE、CDの他に、ベースラインの方法として一次線形補間、二次線形補間による結果も示す。最後に、結果4.4では欠損数を増やした場合の評価、つまりNM=2、NM=3とした場合の各手法の評価値を示す。

4 結果

4.1 NNモデル構築時の学習曲線

図4にWBCの欠損値を推定するAEとCDの単系列モデルにおける、モデル構築時の学習曲線を示す。AE、CD共に1500エポックの時点で誤差の改善はほぼプラトーに達していた。CDはAEに比べて学習の進みが若干早く、また学習セットとテストセットの曲線がほぼ重なっていることから、モデルのフィッティングと汎化性能はCDが良い傾向にあると思われた。AE、CD共にCRを増やすに従って曲線が上方にシフトするが、これはマスクした値が増えるにつれて平均誤差が上昇するためであり、1500エポック終了の時点で最も良い精度を示したのはCR=0%、つまり学習時に一切マスクをかけないものであった。紙面の都合上掲載を省略したが、WBC以外の系列、また多系列のモデルにおいても同様の傾向であった。

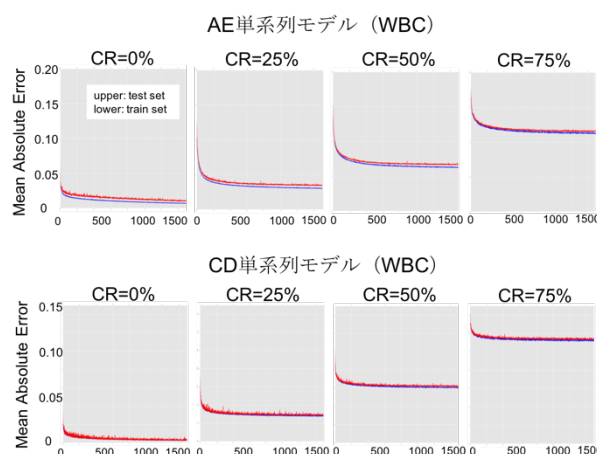


図4 WBCの欠損値を推定する単系列モデルの学習曲線
各サブフィギュアの縦軸は平均絶対誤差を示し、横軸はエポック数を示す。標準化したデータを用いているため、平均絶対誤差は臨床的なスケールとは乖離がある。

4.2 各CRにおけるモデル構築時と評価時の平均絶対誤差の比較

図5に各CRで構築したAEとCDの単系列モデルを用いてNM=1とした場合の評価値を示す。各サブフィギュアにおいて、縦軸は平均絶対誤差、横軸は各CRを意味する。破線はモデル構築時の1500エポック時点の平均絶対誤差を示しており、前述の通りCR=0%のモデルが最も良い値を示すが、実線で示す評価時の平均絶対誤差はCR=0%のモデルが最も高く、CR=25%、50%、75%とした場合には大きな差は見られないものの、CR=50%が最も誤差が低かった。多系列のモデルも同様の傾向であった。この結果から、CR=50%で構築したモデルを用いることとし、以後の評価結果を示す。

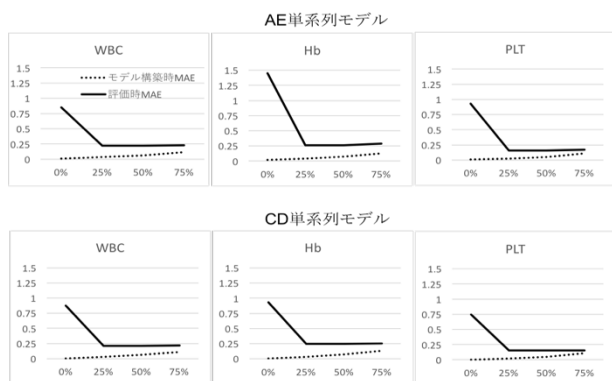


図5 各CRのモデル構築時と評価時の平均絶対誤差

4.3 各種法における欠損値推定の誤差の比較

表2(巻末資料)にNM=1とした場合の各手法における絶対誤差の平均値、中央値、最大値をそれぞれ示す。いずれの指標も値が小さいほど推定精度が良いことを意味する。表中の値はデータ標準化の際に除算した分散の値を掛け、臨床的な解釈が可能な値に戻し掲載した。SSMはベースラインである1次線形補間に比べ誤差の平均値と中央値が高く、2次線形補間よりは低い傾向を示した。また、SSMの多系列モデルは単系列のモデルに比べ誤差が高い傾向にあり、多系列にすることによる精度の改善を認めなかった。AEとCDはベースラインよりも平均値、中央値、最大値ともに低い値を示した。またCDにおける多系列のモデルは単系列のモデルに比べ誤差が低い傾向にあり、多系列にすることによる精度の改善を認めた。全手法の比較では、平均値と中央値に関してはCDが最もよい推定精度を示したが、最大値に関してはその限りではなかった。

4.4 欠損数を増やした場合の各手法の比較

表3(巻末資料)にNMを増やした場合の各手法における絶対誤差の平均値を示す。いずれの手法においてもNMを増やすにつれて誤差の推定精度は悪くなるが、手法間の精度の逆転は見られず、NM=3においても多系列のCDが最も良い値を示した。

5 考察

一般病棟の入院症例に適用することを想定し、日の単位で生じる検体検査結果の欠損を精度よく推定する方法として、SSMとNNによる方法を比較した。ベースラインとして提示した1次線形補間の方法と比べ、SSMは欠損値の推定精度が悪く、NNを用いた方法は推定精度が良かった。また、NNを用いた方法間の比較では、全結合のネットワークであるAEよりも畳み込みフィルタを使ったCDの精度が良かった。以下、SSMとNNを用いた方法の妥当性と結果を検討する。

5.1 SSMの適用妥当性と結果の解釈

検体検査は人体の内部的な状態を血液中の細胞数や化学成分の量によって観察するものであるため、この時系列変化のモデル化にSSMを用いることは妥当と考える。SSMは観測値が欠損する場合にも遷移する状態の値から観測値を推定できるため、欠損値を多く有する医療検査データの将来予測にも用いられる⁷⁾。本研究は欠損値の推定そのものにSSMを利用したが、ベースラインとして提示した1次線形補間よりも精度が悪く、2次線形補間よりも精度が良い傾向にあった。また、実験対象のWBC、Hb、PLTは発生の起源を共にする

ため状態間に相関があると考えられ、多系列の情報を使うことで推定精度が向上すると期待されたが、今回の実験では単系列のモデルよりも精度が悪化する傾向にあった。

5.2 NNモデルの適用妥当性と結果の解釈

NNを用いた方法は、入力データの一部をマスクした上で順方向の推論を行い、マスクした値を復元するようにネットワークを学習することで欠損値を推定した。結果4.2で示したように、CR=0%のモデルは構築時の誤差は最も良いが、評価時の誤差は最も悪かった。この結果は当然であり、CR=0%のモデルはマスクした値を復元するように学習されていないためである。一方、結果4.3で示したように、CR=50%として学習したAEとCDの誤差の推定精度はベースラインとSSMの方法に比べて良い結果を示した。このことから、マスクした値を復元するように学習したモデルが、欠損値の推定に応用できるというコンセプトの有効性が確認できた。対象のデータセットには真の欠損が含まれており、この値に対する推定精度は評価することが出来ないが、本実験で行った評価は、真の欠損はそのままとして、ランダムに人工的に発生させた欠損を対象として評価するものであり、ネットワークが推論する過程においては2種類の欠損に差は無いため、真の欠損もほぼ同等の精度で値が推定されていると考えられる。また、結果4.4では人工的に発生させた欠損の数を増やした場合でも手法間で精度の逆転はみられず、評価結果が一定の頑健性を持つことを示した。

NNを用いた方法間の比較に関して、全ての検査項目でCDはAEよりも高いモデルのフィッティングを示した。また、CDは多系列モデルにすることの効果も認めたが、AEは認めなかった。この原因として、AEは多系列のモデルとするためにパラメータ(結合重み)の数を増やす必要があるため、欠損値推定に利用できる情報が増えることの効果が、過学習による汎化性能の低下という逆効果に相殺された可能性がある。一方、CDの多系列モデルへの拡張は、入力チャネルを増やすのみで、パラメータを増やす必要がなく、欠損値推定に利用できる情報が増えることの効果が見られたと考えられた。ただし、推定精度の改善はわずかであり、この原因としてそもそもWBC、Hb、PLTの間に十分な相関がない可能性も考えられた。また、今回の実験ではプーリング処理を省いて行ったため、これを加えた場合の検討も必要である。

NNを用いた手法の欠点として、最適なCRをあらかじめ決めることが出来ない点があげられる。本実験では結果4.2の検討からCR=50%として構築したモデルを採用したが、最適なCRは対象データがどの程度欠損を有するかに応じて変わり、欠損の頻度が高いデータセットにおいてはCRの値を高く、頻度が低いものはCRの値を低く設定する必要があると思われる。実際に発生する検査データの欠損の頻度は施設によって様々であるため、最適なCRを都度検討する必要がある。もう一つの欠点として、AEとCDは入力データの要素数が固定されるため、より長い期間の欠損を補完するためには、入力データの要素数を変更してモデルを構築しなおすか、複数のウィンドウに分けて補間を繰り返す必要が挙げられる。後者の方法はウィンドウの境界で系列の情報が途切れてしまい推定精度に不利に働くと思われるため、ウィンドウを重ねてその間の平均値を取るなどの工夫が必要となる。NNにおいて可変長の入力を扱うRecurrent Neural Networkを使った検査時系列の欠損値推定も報告されており⁸⁾、このような方法との比較も検討する必要がある。先行研究に関して、Autoencoderを使って欠損値を補完した上で疾患の予後を予

測する報告はみられたが⁹⁾¹⁰⁾¹¹⁾、ConvolutionとDeconvolutionを時系列の欠損値補間に適用した報告は著者の調べる限りみられなかった。

5.3 臨床的な解釈を加えた検討

実験においてベースラインとして使用した1次線形補間は、AEとCDに比べてもそれほど悪い精度ではなかった。NM=3における1次線形補間とCD多系列モデルとの平均絶対誤差を比較すると、その差はWBCで210/uL、Hbで0.06g/dL、PLTで3000/uLであった。一概には言えないが、推定された値を人間がみて判断する限りは、この差は大きな影響を与えないと思われる。しかしながら、大規模にデータ収集して精密に解析を行うような用途では結果に影響を与えるかもしれない。また、いずれのモデルを用いた場合でも平均値と中央値に乖離があり、いくつかの大きくはずれた値が平均値を上昇させていた。大きくはずれた値が存在することは、推定された欠損値を利用するにあたり念頭に入れておく必要がある。

6. 結語

本研究は日の単位で生じる検査値の欠損を精度よく推定することを目的とし、SSMとNNによる方法を比較した。ベースラインとして提示した1次線形補間の方法と比べSSMは推定精度が悪く、NNを用いた方法は推定精度が良かった。NNを用いた方法間の比較では、ConvolutionとDeconvolutionを使ったモデルはAutoencoderを使ったモデルよりも推定精度が良く、また多系列のモデルにすることの効果も認めた。

7. 倫理的配慮

本研究の実施は東京大学大学院医学系研究科・医学部倫理委員会により承認を得ている(承認番号 10791)

8. 謝辞

本研究は JST さきがけ JPMJPR1654 と MEXT 科研費

16K09161 の助成を受けたものである。

参考文献

- 1) 岩波データサイエンス刊行委員会. 岩波データサイエンス Vol.6.
- 2) 野村俊一. カルマンフィルタ-Rを使った時系列予測と状態空間モデル-. 統計学 One Point2 巻.
- 3) Hinton GE, Salakhutdinov RR. Reducing the Dimensionality of Data with Neural Networks. Science, 2006 Jul 28;313(5786):504-7.
- 4) Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly Detection: A Survey, ACM Computing Surveys, Vol.41(3), Article 15, July 2009.
- 5) P.Vincent, H.Larochelle Y.Bengio and P.A.Manzagol, Extracting and Composing Robust Features with Denoising Autoencoders, Proceedings of the Twenty-fifth International Conference on Machine Learning, pages 1096-1103, ACM, 2008.
- 6) Zeiler MD, Krishnan D, Taylor GW, Fergus R. Deconvolutional networks. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2010. p. 2528-2535.
- 7) Hatakeyama Y, Kataoka H, Nakajima N, Watabe T, Fujimoto S, Okuhara Y. Prediction model for glucose metabolism based on lipid metabolism. Methods Inf Med. 2014;53(5):357-63.
- 8) Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. arXiv:1606.01865.
- 9) Beaulieu-Jones BK, Moore JH. Missing data imputation in the electronic health record using deeply learned autoencoders. Pac Symp Biocomput. 2016;22:207-218.
- 10) Lovedeep Gondara, Ke Wang. Multiple Imputation Using Deep Denoising Autoencoders. arXiv:1705.02737.
- 11) Beaulieu-Jones BK, Greene CS. Semi-supervised learning of the electronic health record for phenotype stratification. J Biomed Inform. 2016 Dec;64:168-178.

巻末資料

表2 各手法における推定欠損値の絶対誤差(NM=1)

	WBC (×10 ³ /uL)			Hb (g/dL)			PLT (×10 ⁴ /uL)		
	平均値	中央値	最大値	平均値	中央値	最大値	平均値	中央値	最大値
線形補間 (1次)	1.51	0.86	236.45	0.57	0.40	10.09	2.23	1.40	74.30
線形補間 (2次)	1.84	1.01	188.46	0.71	0.50	19.06	2.52	1.58	98.00
SSM (単)	1.61	1.00	200.92	0.58	0.42	14.65	2.39	1.50	97.55
SSM (多)	1.74	1.04	237.36	0.59	0.42	16.14	2.39	1.46	104.29
AE (単)	1.44	0.85	170.61	0.56	0.42	8.32	2.13	1.42	71.90
AE (多)	1.47	0.87	190.47	0.55	0.41	8.23	2.11	1.45	80.61
CD (単)	1.38	0.81	244.90	0.54	0.40	8.48	2.06	1.33	64.60
CD (多)	1.35	0.79	242.18	0.53	0.39	8.87	2.03	1.34	70.97

表3 欠損数を増やした場合の各手法間の比較(絶対誤差の平均値)

	NM=1			NM=2			NM=3		
	WBC	Hb	PLT	WBC	Hb	PLT	WBC	Hb	PLT
線形補間 (1次)	1.51	0.57	2.23	1.60	0.60	2.44	1.65	0.62	2.62
線形補間 (2次)	1.84	0.71	2.52	1.90	0.72	2.65	N/A	N/A	N/A
SSM (単)	1.61	0.58	2.39	1.72	0.61	2.55	1.78	0.64	2.66
SSM (多)	1.74	0.59	2.39	1.87	0.62	2.56	1.92	0.65	2.66
AE (単)	1.44	0.56	2.13	1.47	0.57	2.24	1.51	0.59	2.37
AE (多)	1.47	0.55	2.11	1.49	0.56	2.28	1.52	0.58	2.40
CD (単)	1.38	0.54	2.06	1.41	0.56	2.19	1.46	0.58	2.35
CD (多)	1.35	0.53	2.03	1.38	0.54	2.17	1.44	0.56	2.32