経過記録情報を用いた深層学習による欠損値補間したHbA1c予測モデルの構築

畠山　豊、兵頭　勇己、奥原　義保
高知大学　医学部附属　医学情報センター

Prediction model of HbA1c with interpolated missing values based on deep learning by text data

Yutaka Hatakeyma, Yuki Hyohdoh, Yoshiyasu Okuhara
Center of Medical Information Science, Kochi Medical School, Kochi University

抄録: 急激なHbA1c値変動を予測する状態空間モデルが長期間の変動を追随可能にするため、経過記録に基づいて欠損値補間を行いながら予測計算を行う手法を提案する。モデルパラメータとして、HbA1c値とその変動量を定義し、観測値が得られた時刻では値に基づきパラメータを修正し、変動量が得られない時刻では対象時刻の経過記録から変動量に関する2クラス識別を行う事前学習済みの深層学習モデル出力値に基づきパラメータ修正を行う。高知大学医学部附属病院の患者データに提案手法を適用した結果、初回HbA1c検査から1200日後のHbA1c値に対するRMSEが0.30となり補間を行わない手法に比べ約半分に減少した。状態空間モデルは欠損値が存在しても時系列変動を扱うことができるが、経過記録に基づく観測データ生成を行うことで予測精度向上が実現できるため、提案手法は実診療データの予測処理に有用である。

キーワード: 予測モデル、粒子型フィルタ、深層学習、テキストマイニング

1.　はじめに

糖尿病の早期介入を行うためHbA1c値の時系列変動予測モデルは重要である。発症直前の急激な変化を予測するため、非線形モデルによる手法[1]が提案されている。このような急激な変化を精度よく予測するためには、検査値だけでなく一定期間内の個体内変動量が有効であるが、実診療ではHbA1cは頻回に実施されず、変動量が欠損値となり、その結果長期予測では予測誤差が蓄積してしまう。一方、受診時に記載している経過記録には、医師が判断した患者状態や処方などの治療行為が記載されているため、この記載パターンから糖尿病の進行度合いが判断できる可能性がある。

本論文では、HbA1c値と一定期間内の変動量を状態モデルとした状態空間モデルによる予測手法を提案する。変動量が欠損している場合、経過記録から深層学習により変動量クラスを推定し観測データとして用いることで長期間の変動予測精度の向上を実現する。高知大学医学部附属病院の患者データに適用し、長期間の時系列変動に提案手法が適切に追随可能であるかどうかを評価する。

2.　方法

1)　状態空間モデルに基づく予測モデル

各時刻におけるHbA1c値を予測するモデルを状態空間モデルによって記述する。その概要をFig.1に示す。状態モデルでは真のHbA1c値($Data_s$)と真のHbA1c変動量($\Delta Data_s$)を記述し、各状態モデル値に観測ノイズを加えたデータを観測モデルとして定義する。実際の計算はパーティクルフィルタで行う。

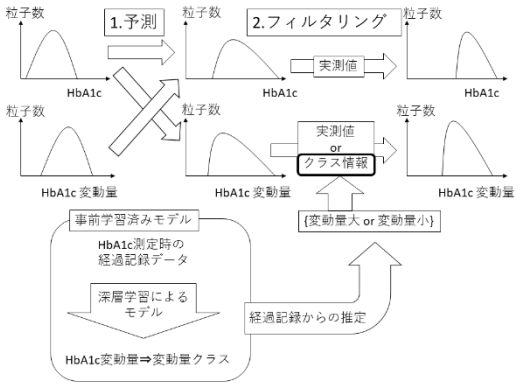

Fig.1　予測モデルの概要

時刻tの$Data_s(t)$及び$\Delta Data_s(t)$から次時刻t+1の値を以下の式で予測する

$$Data_s(t+1) = Data_s(t) + f\left(\Delta Data_s(t)\right) + N_1$$

$$\Delta Data_s(t+1) = g\big(Data_s(t)\big) + N_2$$

関数f,gは、それぞれ線形関数、シグモイド関数であり、N_1は分散が$\Delta Data_s(t)$に依存する正規乱数N_2は分散が固定値の正規乱数である。

各粒子の予測値と実際の測定値との尤度に基づきフィルタリングを実施する。HbA1c が1回だけ測定している場合、経過記録から変動量クラスを構築済みモデルから算出し、尤度を計算する。1回も測定していない場合は次時刻予測を継続する。尤度は正規分布の確率密度に基づき定義する。

2) 経過記録から変動量のクラス識別する深層学習モデル

自然言語を対象とする深層学習モデルである BERT モデルの事前学習済みモデル[2]を利用して、入力を経過記録とし、出力を HbA1c 変動量についての 2 クラスとした識別モデルを fine tuning によって構築する。

3) 実験対象データ

2009 年から 2019 年の間、高知大学医学部附属病院において HbA1c を測定した患者を対象とする。1単位期間を 120 日とする。識別モデルの学習データとして経過記録の文字数が 300 から 700 であり、期間中複数回測定している 1940 件を用いて構築し、277 件のテストデータで評価する。クラス識別の閾値を 0.3 とする。予測モデルの対象として、初回測定から 10 期間の対象期間中、最終期間と対象期間中 7 期間1回以上測定を行っている 192 人のデータを用い予測計算を実施した。

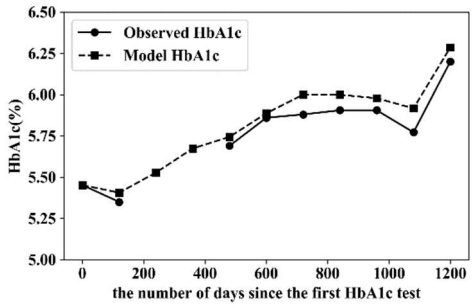

Fig.2　1 患者データにおける HbA1c 提案モデルフィルタリング値
（実測値:実線、フィルタリング値:点線）

3. 結果

構築した識別モデルのテストデータに対する精度は 0.74 となった。ある1患者における HbA1c 値及びフィルタリング後の分布重心を Fig.2 に示す。

最終期間におけるモデル予測分布及びフィルタリング分布における重心値と実測値との二乗平均平方根誤差(RMSE)はそれぞれ 0.30 と 0.14 となった。経過記録を利用しない場合、つまり、変動量が観測できた場合のみフィルタリング処理を行う場合での最終期間における予測及びフィルタリング後の重心値の RMSE は 0.59 と 0.17 となった。

4. 考察

深層学習を用いても、識別モデルが必ずしも高精度ではない理由として、糖代謝やその治療に関わる記載が少ない経過記録が一部存在するためと考える。しかし、観測誤差が存在する前提で状態空間モデルや尤度を定義しているため、時系列変動に適切に追従できていることをフィルタリング後の RMSE 値は示していると考える。

クラス分類情報を観測データとした場合における予測分布の RMSE が約半分に減少したことは、経過記録で観測データを適切補間したことで予測精度向上が実現したことを示していると考える。

実診療データは欠損値が多く存在しているため、欠損でも計算可能な状態空間モデルは有効であるが、予測精度が低下するため、本論文で示した経過記録データから観測データを補間する手法は時系列データモデルに有効であると考える。

限界として、予測対象患者の経過記録文字数が対象外のデータも含まれているため、識別モデルの出力精度が低い可能性がある。

5. おわりに

経過記録から変動量クラスを識別する学習モデルにより欠損値補間する状態空間モデルは時系列変動を適切に予測可能なことを確認できた。

参 考 文 献

[1] Hatakeyama Y, Kataoka H, Nakajima N, et al: Prediction model for glucose metabolism based on lipid metabolism. Methods Inf Med. 53(5):357-63. 2014.

[2] BERT Pretrained model Trained On Japanese Wikipedia Articles https://github.com/yoheikikuta/bert-japanese (2020/02/01)