機械学習による糖尿病リスク予測ツール（第２版）の開発

要約

職域多施設研究（J-ECOH スタディ）において、解析対象者の 3 分の 2 にあたる 29,265 件とその後の糖尿病発症データを用いて、糖尿病の発症を予測するニューラルネットワークのモデルを構築し、その予測精度を残りの 14,597 件のデータで検証しました。予測精度の指標である ROC 曲線下面積（AUC）は、血液検査データを含まない変数のみで作成したモデルでは 0.759、さらに空腹時血糖・ヘモグロビン A1c などのデータを追加して作成したモデルでは 0.921 で、特に後者では高い予測精度であることが確認されました。

1. 開発に使用したデータ

職域多施設研究（J-ECOH スタディ）で収集された健康診断データにもとづいてツールを開発しました。2010 年度をベースラインとして、2013 年度までの 3 年間において、空腹時血糖 126mg/dL 以上、ヘモグロビン A1c 6.5%以上、糖尿病治療のいずれかの条件に該当した場合を糖尿病発症と定義しました。

以下の条件にあてはまるデータは除外しました。 1)ベースライン時点で年齢が 30 歳未満、または 65 歳以上である、2)心血管疾患、がん、糖尿病の既往がある、3)非空腹状態で血液検査を受けた、4)説明変数のデータが欠損している、5)2013 年度末時点で糖尿病発症が判定できない。6)ベースライン時点でヘモグロビン A1c4.0 未満、空腹時血糖 60 未満、拡張期血圧 40 未満、AST(GOT)200 超、ALT(GPT)200 超、γ-GTP 800 超。その結果、43,862 件のデータが残り、このうち 3 分の 2 にあたる 29,265 件をモデル構築用、3 分の 1 にあたる 14,597 件を検証用としました。

表 1 の 4 つのモデルを構築しました。パターン 1 は血液データを用いない非侵襲性モデル、パターン 2～4 は血液データを用いた侵襲性モデルです。機械学習を行う前に、各項目に対し、次の 3 つのいずれかの方法でデータを変換しました。

表 1. リスク予測に用いた項目

変数名	型	パターン 1	2	3	4
目的変数					
糖尿病罹患	カテゴリ	○	○	○	○
説明変数					
性別 (男性：1，女性：0)	カテゴリ	○	○	○	○
年齢	数量	○	○	○	○
身長 [cm]	数量	○	○	○	○
BMI [kg/m^2]	数量	○	○	○	○
腹囲 [cm]	数量	○	○	○	○
喫煙 (現在)	カテゴリ	○	○	○	○
ヘモグロビン A1c [%]	数量	-	-	○	○
空腹時血糖 [mg/dl]	数量	-	○	-	○
LDL コレステロール [mg/dl]	数量	-	○	○	○
HDL コレステロール [mg/dl]	数量	-	○	○	○
中性脂肪 [mg/dl]	数量	-	○	○	○
脂質異常症治療	カテゴリ	○	○	○	○
収縮期血圧 [mmHg]	数量	○	○	○	○
拡張期血圧 [mmHg]	数量	○	○	○	○
高血圧治療	カテゴリ	○	○	○	○
AST(GOT) [U/L]	数量	-	○	○	○
ALT(GPT) [U/L]	数量	-	○	○	○
γ - GTP [/L]	数量	-	○	○	○
ヘモグロビン [g/dl]	数量	-	○	○	○

1)変換なし：値の変換を行わない。カテゴリデータに対して使用。2)正規化：値の範囲を 0～1 にする。3)対数正規化：対数をとった後、値の範囲を 0～1 にする。

2. 予測モデルの構築

開発には機械学習法のひとつであるニューラルネットワークを用いました。ニューラルネットワークのモデルの学習には DeepLearning 用フレームワークである Chainer を使用しました。入力層のユニット数はパターンにより異なり、出力層のユニット数は 1 です。出力層の入力値に対して sigmoid 関数により範囲を(0,1)に変換しました。この値が大きいほど、糖尿病を発症する可能性が高いデータと解釈できます。モデルの学習時、各中間層で Dropout を行い、ミニバッチ数を 50、エポック数を 50 と設定しました。また、発症者件数と非発症者件数の不均衡対応のため、オーバーサンプリング手法[1]を使用しました。

モデル構築用データを用いて、パラメータを探索しました。探索するパラメータはハイパーオプト[2]により選ばれ、その評価には k 分割交差検証法

[1] Chawla, N. V. et al.: SMOTE: Synthetic Minority Over-sampling Technique, J. Artif. Int. Res., Vol. 16, No. 1, pp. 321–357 (2002)

[2] Bergstra, J., Yamins, D., and Cox, D. D.: Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms (2013)

[3]を使用しました（本開発では k=5）。計 500 通りのパラメータを試した結果、最も評価値の高かったパラメータを採用しました。

モデル学習時にデータの調整を行っているため、リスク算出にあたって再調整用のロジスティック回帰式を作成しました。ある方の健康診断データをモデルに入力すると、その出力値が計算され、その値をロジスティック回帰式に代入します。その式の値をその方の発症リスクとしてお返しします。

3. 予測モデルの精度評価

構築したモデルの予測精度を評価するため、検証用データを用いて各人の出力値を計算しました。発症と判定する閾値を変え、そのときの偽陽性率と真陽性率を計算・プロットして、Receiver Operating Characteristic curve（以下、ROC 曲線）を描き、ROC 曲線下面積（以下、AUC）を計算しました。AUC は 0～1 の範囲をとるモデルの性能を表す指標であり、値が高いほど性能がよいとされます。

各パターンのモデルに対応した ROC 曲線を**図**に、また AUC の値を**表 2** に示します。AUC はパターン 1 が最も低く、パターン 4 が最も高い値となりました。南里らは今回と同じ研究に基づいてロジスティック回帰により予測モデルを開発していますが[4]、その AUC は非侵襲モデルで 0.734、空腹時血糖を含む侵襲モデルで 0.835、ヘモグロビン A1c を含む侵襲モデルで 0.819、空腹時血糖とヘモグロビン A1c 両方を含む侵襲モデルで 0.882 と報告されています。今回開発したモデルは使用している変数の数や取り扱いが若干異なるため厳密な比較はできませんが、それぞれ 0.025、0.056、0.075、0.039 の向上が見られます。

すなわち、ニューラルネットワークを用いて開発した本モデルは、統計理論に基づいて開発した従来のモデルより、特に血液データを用いた場合において高い予測能を持っているといえます。

図. 4 つの予測パターンの ROC 曲線

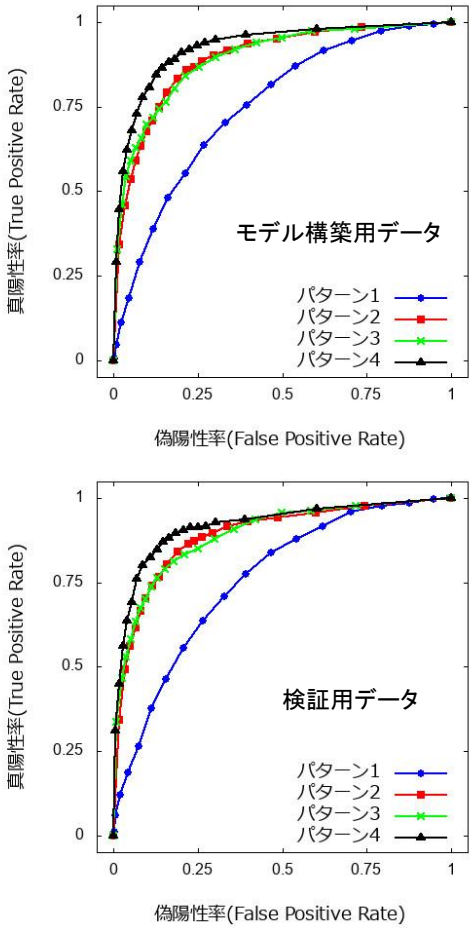

表 2. 各パターンの AUC

モデル パターン	構築用データ AUC	検証用データ AUC
パターン 1	0.753	0.759
パターン 2	0.894	0.891
パターン 3	0.894	0.894
パターン 4	0.928	0.921

4. 今回の改定点

データに関して

- 対象年齢を 30～64 歳に広げた。
- 説明変数の異常値による除外基準を追加した。
- 説明変数の変換方法を項目ごと決定した。

[3] モデル構築用データを k 個に分割し、その中の k-1 個のデータで学習し、残りの 1 個をテストデータとし評価値を算出する方法。テストデータを変更して、合計 k 個の評価値の算出後、その平均値をそのパラメータの評価とする。

[4] Nanri, A. et al.: Development of Risk Score for Predicting 3-Year Incidence of Type 2 Diabetes: Japan Epidemiology Collaboration on Occupational Health Study, PLOS ONE, Vol. 10, pp. 1–16 (2015)

発症リスクの算出方法に関して
機械学習の出力値と実際の罹患状況によりロジスティック回帰式を作成し、この式により発症リスクを算出した（連続的な発症リスク）。