ポスター

ポスター8
医療データ分析４（データ・テキストマイニング）

2018年11月23日(金) 17:00 〜 18:00 K会場(ポスター、HyperDemo) (2F 多目的ホール)

[2-K-4-6] 時系列成分を考慮した線形回帰による血糖コントロール予測

○由井 俊太郎, 大崎 高伸, 伴 秀行 （（株）日立製作所　研究開発グループ）

健康保険組合では，健診やレセプトなどのデータ分析に基づき，生活習慣病予防や重症化予防のための効果的な保健事業を行うデータヘルスを推進している。特に糖尿病は，合併症が重篤になるため，発症予防だけでなく，糖尿病発症者の重症化予防も重要である。このため健康保険組合では，現状の血糖コントロール状態に基づき，医療機関未受診者への受診勧奨などの重症化予防対策を進めているが，現状だけでなく，糖尿病発症者の将来の血糖コントロール状態を予測できれば，事前にコントロール状態が悪くなる可能性が高い対象者についても対策が可能になると考えられる。そこで，糖尿病発症者の効果的な重症化予防支援を目的に，健診とレセプトのデータを活用し，血糖コントロール状態を高精度に予測する方法を提案する。通常，診療行為の実施有無に基づいて，線形回帰などの手法を用いて，将来の血糖コントロール状態を予測する方法が用いられる。しかし，健診データやレセプトデータには，検査値を取得した時期や診療行為の実施した時期も含まれるため，実施有無に加えて，時系列の情報も考慮することで，さらに高精度に血糖コントロール状態を予測できるのではないかと考えた。そこで提案手法では，説明変数のデータセットを生成する際に，時系列成分を考慮してデータセットを生成することとした。提案手法の有効性を示すため，糖尿病発症者8820人の2013年から2016年の健診とレセプトデータを用いて，モデル精度を表す AUCを交差検証して従来手法と比較した。なお，説明変数は2013年の検査値と2013年から2015年の診療行為から生成し，目的変数は2016年の血糖コントロール状態とした。その結果，従来の実施有無のみを使った手法の AUCが0.784であったのに対して，提案手法の AUCは0.805と改善した。時系列成分を考慮した提案手法のほうが AUCは高くなり，血糖コントロール状態をより高精度に予測できる可能性を見出した。

©一般社団法人 日本医療情報学会

時系列成分を考慮した線形回帰による血糖コントロール予測

由井俊太郎[*1]、大崎高伸[*2]、伴秀行[*3]
*1(株)日立製作所 研究開発グループ

Time-series Information Compressing Linear Model for Blood Glucose Control Prediction

Shuntaro Yui [*1] , Takanobu Osaki[*1] , Hideyuki Ban[*1]
*1 Center for Technology Innovation, R&D, Hitachi Ltd.

Japanese health insurance societies are now conducting “Data Health”, which is a data-driven approach not only for cost saving but also for improvement of the quality of life. In the Data Health, diabetes prevention is one of their focus areas as the cost burden of diabetes is enormous due to its co-morbidity and complication. From a data-driven perspective, risk subject identification could help to achieve efficient prevention service, however, a lot of insurance societies are able to access to limited data such as claim and health check data, including data sparseness which may cause difficulty of data leveraging. In order to tackle this problem, we proposed a high accuracy prediction method for blood glucose control. Proposed method introduces time-series information compressing in order to leverage massive data while to reduce data sparseness. Experimental results using real data of subjects newly diagnosed with diabetes showed that Area Under the Curve (AUC) of our proposed model has improved from 0.784 to 0.805. It demonstrated that the proposed method could achieve high accuracy of prediction.

Keywords: Data Health, Diabetes Prevention, Prediction Model

1. 緒論、目的

健康保険組合では、生活の質の改善と医療費の適正化を同時に目指す保健事業を進めている。2015 年度からは、健診やレセプトなどのデータを用いた分析に基づき、生活習慣病予防や重症化予防のための効果的な保健事業を行うデータヘルスが推進されている [1)]。特に糖尿病では、心臓疾患や腎症など合併症が重篤になるため、発症予防だけでなく、糖尿病発症者の重症化予防も重要な課題の一つである。このため健康保険組合では、現状の血糖コントロール状態に基づき、医療機関未受診者への受診勧奨などの重症化予防対策を進めている。しかしながら、現状の状態だけでなく、糖尿病発症者の将来の血糖コントロール状態を予測できれば、コントロール状態が将来悪くなる可能性が高い対象者についても事前に早期対策が可能になると考えられる。

そこで、糖尿病発症者の効果的な重症化予防支援を目的に、健診とレセプトのデータを活用し、血糖コントロール状態を高精度に予測する方法を提案する。

2. 方法

従来の血糖コントロール状態を予測する手法として、診療行為の実施有無を用いた線形回帰がある。しかし、健診データやレセプトデータには、検査値を取得した時期や診療行為の実施した時期も含まれるため、実施有無に加えて、時系列の情報も考慮することで、さらに高精度に血糖コントロール状態を予測できるのではないかと考えた。そこで提案手法では、説明変数のデータセットを生成する際に、時系列成分を考慮してデータセットを生成することとした。

以下、本手法の流れを示す。

<u>Step1: Data extraction</u>

まず、オリジナルのデータソース（健診データベース、レセプトデータベース）からデータを抽出する。健診データベースからは、主に患者属性と検査値の時系列変化を抽出する。レセプトデータベースからは、病気、処方、検査の実施有無を、実施した時期の情報と共に抽出する。詳細な抽出項目を表 1 に示す。

<u>Step2: Explanatory variable creation</u>

時系列成分を考慮した説明変数のデータセットを生成する。まず始めに、患者 i ごとに糖尿病を発症した月 M(i)を抽出する。次に、表 1 で定義したレセプトデータから抽出した診療行為群 A = {A1, A2, …, AJ} (e.g. A1 = Diabetes, A2 =Sulfonylureas)に対して、糖尿病発症月からの相対月 Rij(t)を患者 i 毎に抽出する。次に、患者 i の診療行為 Aj の実施状況を集約した Xij を生成し、時系列成分を考慮した診療行為に関する説明変数のデータセットを生成する。最後に、診療行為に関する説明変数のデータセット Xij に、発症イベント時の検査値を患者 i 毎に結合する。

<u>Step3: Outcome prediction based on the linear model</u>

Step2 で生成したデータを説明変数、血糖コントロール状態を目的変数として、線形回帰モデルを生成する。今回、血糖コントロール状態は、糖尿病の重要な血糖コントロール指標である HbA1c(NGSP)を用い、コントロール基準は、特定健康診査の受診勧奨判定値である 6.5%未満を設定した [2)]。また、特徴選択手法として、線形回帰をベースとする下記3手法から、交差検証の中で最も Area Under the Curve (AUC)の高いモデルを選択する。

- Ridge Regression [3)]
- Least Absolute Shrinkage and Selection Operator [3)]
- Elastic Net [4)]

3. 結果、考察

提案手法の有効性を示すため、糖尿病発症者 8820 人の 2013 年から 2016 年の健診とレセプトデータを用いて、モデル精度を表す AUC を交差検証して従来手法と比較した。なお、説明変数は 2013 年の検査値と 2013 年から 2015 年の診療行

為から生成し、目的変数は 2016 年の血糖コントロール状態とした。

本手法の予測精度を測定するため、5X2 cross validation を行い、モデル精度を表す AUC を用いて従来手法と比較した。ここで従来手法とは、診療行為の回数を患者別に集計したデータを説明変数として、2. Step3 に示す 3 種類の特徴選択手法に入力したものとする。

図 1 に、AUC の比較結果を示す。従来手法の AUC が 0.784 に対して、提案手法の AUC は 0.805 と高いことが分かる。また、図2と図3に、従来手法と提案手法の ROC 曲線をそれぞれ示す。これらのことから、提案手法の汎化性は従来手法と変わらず、過学習にも対応できることが分かる。

表 1 Random variables in this study

Data Source	Data Category	Variables
Health check	Demographic	Gender, Age
	Test result measurement	Body Mass Index (BMI), HbA1c, Low-Density Lipoprotein (LDL)
Claim data	Diagnosis	Diabetes, Hypertension, Hyperlipidemia, Retinopathy, Renal, Chronic kidney disease, Nervous, Heart infarction, Stroke, Cerebral bleed, Cardiac hypertrophy, Atrial fibrillation, Heart failure, Aortic disease , Arteriosclerosis, Gestational diabetes
	Medication	Sulfonylureas, Insulin, Glucosidase inhibitor, Biguanide, Thiazolidinedione, DPP4 inhibitor, SGLT2 inhibitor, Compounding agent, Glucagon like peptide-1receptor agonist, Hypotensive agent, Hypolipidemic agent
	Test	HbA1c, Blood glucose, Albumin, 1-5-anhydroglucitol, Glucose tolerance trend test, Glucose tolerance test, Blood glucose self-test, Total cholesterol, Triglycerides, High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), Lifestyle intervention (Diabetes) , Lifestyle intervention (Hypertension), Lifestyle intervention (Hyperlipidemia), Dialysis, Food habit intervention

4. 結論

従来の実施有無のみを使った手法の AUC が 0.784 であったのに対して、提案手法の AUC は 0.805 と改善した。時系列成分を考慮した提案手法のほうが AUC は高くなり、血糖コントロール状態をより高精度に予測できる可能性を見出した。

参考文献

1) データヘルス計画作成の手引き. 厚生労働省。Available at [https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000061273.html]

2) 標準的な健診・保健指導プログラム【平成 30 年度版】. 厚生労働省。Available at [https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000194155.html]

3) Y. Saeys et al, "A review of feature selection techniques in bioinformatics. Bioinformatics," 23(19):2507-2517, 2007.

4) Zou H et al, "Regularization and variable selection via the elastic net," J. Royal. Stat. Soc. B. ;67(2):301-320, 2005

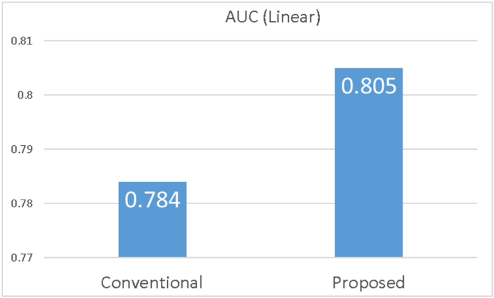

図 1 AUC comparison

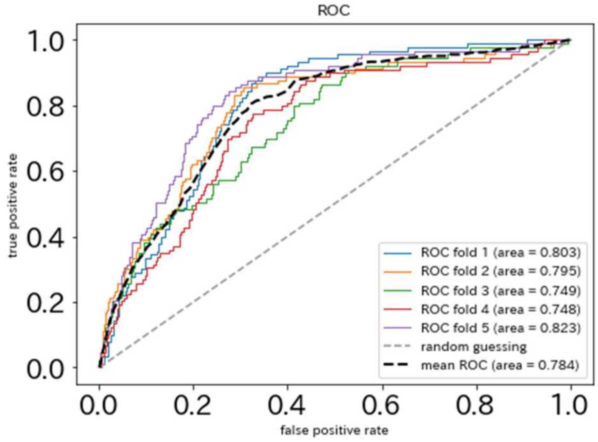

図2 ROC curve (Conventional)

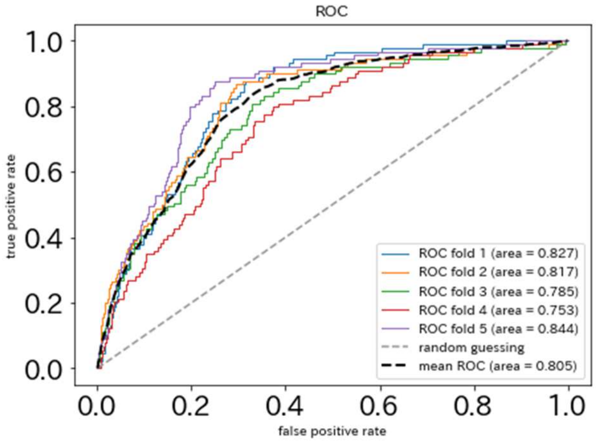

図3 ROC curve (Proposed)