

レセプトデータを用いた生活習慣病の発症予測

Prediction of the Onset of Lifestyle-related Diseases Using Health Insurance Claims Data

八重樫 文絵 ^{*1}

Fumie Yaegashi

荒木 雅弘 ^{*1}

Masahiro Araki

岡 夏樹 ^{*1}

Natsuki Oka

新谷 元司 ^{*2}

Motoshi Shintani

吉川 昌孝 ^{*3}

Masataka Yoshikawa

^{*1}京都工芸繊維大学

Kyoto Institute of Technology

^{*2}SG ホールディングスグループ健康保険組合

SG Holdings Group Health Insurance Association

^{*3}日本システム技術株式会社

Japan System Techniques Co., Ltd.

This paper proposes a system which predicts the onset of lifestyle-related diseases using health insurance claims data. In the transportation industry, they try to reconsider health management to take measures against drivers' overwork these days. Previous studies used representation learning for predicting some diseases. Similarly, we regard this issue as a text classification problem in natural language processing and try to make a model which helps drivers' health management. We transformed the health insurance claims into a fixed-length vector and predicted lifestyle-related diseases with UnderSampling and Bagging. As a result, our model achieved 0.75 in the recall of positives. We're sure that the significance of applying natural language processing to health insurance claims data was shown in this study.

1. はじめに

近年、医療情報の電子化により、健康診断結果やレセプトデータを用いた疾病発症リスクの予測や、それに伴う医療費の削減が様々な業界で求められている。特に運送業界では、昨今のインターネット通販の急速な普及による宅配件数の増加に伴うドライバーの過重労働は社会的問題となっており、運転中の事故率は年々増加しつつある[国土交通省 17]。ドライバーの健康管理対策を見直すために新たな対策の考案が急務とされ、医療情報を基にした生活習慣病等の発症予測モデルの開発が期待されている。

発症予測を目的とした先行研究には、血液中の 1130 種類のタンパク質から心疾患に関連する 9 種類のタンパク質を特定し、心疾患リスクスコアの推定法を提案した研究[Ganz 16]や、健診結果に基づいて心筋梗塞や脳梗塞の発症確率を予測する研究[Yatsuya 16]がある。これらの研究では、血圧値や採血結果の数値データを主に使用しており、レセプトのような文字列の多いデータには適用できない。そこで、Skip-gram を利用した心不全予測の研究[Choi 17]や、Bag-of-Words を用いた研究[Weng 17]、NN による単語の分散表現を利用した研究[Choi 16][Chen 17][Liu 18][Bo 18]など、医療情報を表現学習として扱っている研究のように、自然言語処理の観点から考えて、本問題を文書分類問題として捉え、可変長データを固定長ベクトル化する方法を試行できないかと考えた。そこで本研究では、特定の運送業者のレセプトデータを用いて生活習慣病に属する複数の疾病名を対象に、今後 1 年以内に発症する可能性のある人を予測することを目的とする。この問題を解決するために我々は、リカレントニューラルネットワークを用いてレセプトデータを固定長ベクトル化し、アンサンブル学習によって識別器を構成する方法を提案する。なお、対象とする疾病名及びそれぞれに対応する関連保健問題の国際統計分類(ICD)10 は表 1 に示す通りである。

表 1: 予測対象の疾病名

疾病名	ICD10
インスリン依存型糖尿病	E10
インスリン非依存型糖尿病	E11
糖尿病	E14
狭心症	I20
急性心筋梗塞	I21, I22
心筋症	I42
不整脈・伝導障害	I44~I49
くも膜下出血	I60, I690
脳内出血	I61, I691
脳梗塞	I63, I693

2. 提案手法

本研究では、特定健康保険組合が持つ特定個人に関する全てのレセプトを時系列順に並べたものを 1 つの文章として捉え、形成したデータを固定長ベクトル化し、最後にアンサンブル学習で識別器を構成する手法を提案する。

2.1 レセプトデータの扱い

レセプトデータとは患者に対する診療行為や調剤の記録であり、この多くが文字列として保存されている。機械学習で利用するために、我々はこれらの文字列を数値化する方法を検討した。ICU 入室患者の死亡リスク予測を目的とした Nori らの研究[Nori 17]では、医療費の発生する全ての介入行為に対する属性を 1 とすることで特徴量に加えて数値化を図っているが、これではレセプトデータの時系列性を考慮できない。そこで本研究では、特徴量に性別、年齢、“外来または入院”、摘要名を選択し、以下の例のように各人毎に時系列順にこれらを並べ、1 つの文章として解釈した。この時、摘要名中の薬名はその薬の一般名に置き換える。一般名とはその薬の主成分のことを指し、このように置換することでジェネリック医薬品に

連絡先: 〒 606-8585 京都府京都市左京区松ヶ崎橋上町 1
 京都工芸繊維大学大学院 工芸科学研究科 情報工学専攻
 インタラクティブ知能研究室, yaeg@ii.is.kit.ac.jp

移行した後の変化や、効能が同じなのに別会社の製品であるという違いを吸収できる。

例：Aさんの場合

レセプト 1:[男性, 40 歳, 外来, 初診, インフルエンザ検査, ロキソニン 100mg]

レセプト 2:[男性, 41 歳, 外来, 再診, 大塚生理食注]

↓

[男性 40 歳 外来 初診 インフルエンザ検査 ロキソプロフェンナトリウム水和物 男性 41 歳 外来 再診 生理食塩水]

2.2 学習データ形成

以下、表 1 に示す疾病名の診断を受けたことがない人を健康者、受けたことがある人を重症者として、それぞれの学習データの形成方法を示す。

まず健康者の場合である。健康者は全員、直近 1 年間のデータを削除する。本稿執筆時点では 2018 年 5 月のデータが最新なので、健康者は 2017 年 5 月以前のデータを使用する。次に重症者の場合、重症疾病名初出から n カ月前以前のデータを使用する。 n を 1 から 12 まで一様に分布させ、1 人につき 12 個のデータを作成してデータの拡張を行う。このように“発症の n カ月前までのデータを使う”ようにすることで、数少ない重症者のデータは発症直前まで捨てることなく利用し、かつ、重症者の情報が常に発症直前までのデータであるという偏りを防ぐことができる。図 1 に健康者と重症者のデータ処理方法を示す。

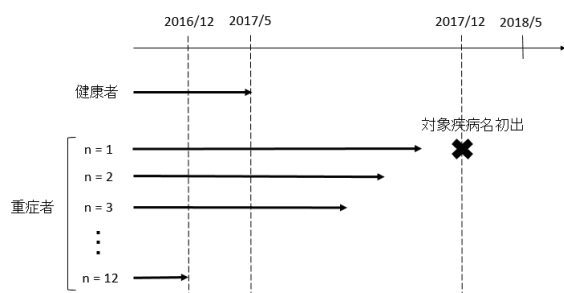


図 1: データの削除

このようにしてできた重症者のデータを図 2 に示すように 12 個に割り振る。ここで、 n 個目の Pot に n カ月前以前のデータばかりが集まるのを防ぐため、人毎に機械的に n を 1 ずつずらして分割していく。

	Pot1	Pot2	Pot3	Pot4	...	Pot12
A	n = 1	n = 2	n = 3	n = 4	...	n = 12
B	n = 2	n = 3	n = 4	n = 5	...	n = 1
C	n = 3	n = 4	n = 5	n = 6	...	n = 2
.			.			
.			.			
.			.			

図 2: データの割り振り

こうして作成された重症者のみのデータ群 12 個に対して、次は図 3 に示すように健康者のデータを付加していく。ここで圧倒的に健康者のデータが多いことに対する対策として UnderSampling を行う。重症者と同じ人数だけ健康者の中からランダムにデータを抽出し、Pot のデータ群に付加していく。さらにデータ量を増やすためにこの作業を 1 つの Pot に対して 10 回行う。これを 12 個全ての Pot に適用し、最終的に 120 個の学習データを作成する。

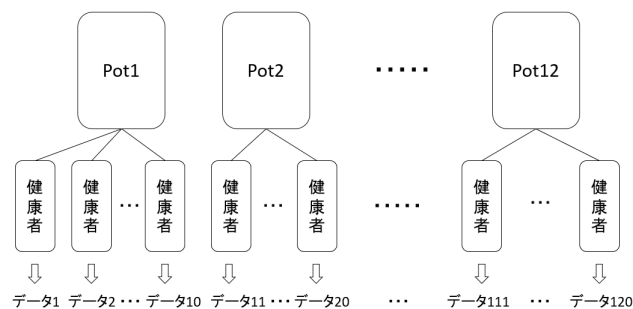


図 3: データの UnderSampling

2.3 固定長ベクトル化

レセプトデータでは、人ごとに通院回数も異なれば一回の診察で行われる診療行為や調剤の数も異なる。つまり形成されたデータは可変長である。この可変長なデータを、図 4 に示すように固定長に変換するために、文書分類問題でよく使用される TF-IDF を利用する方法と、時系列性を加味できる RNN の中間層の出力を取り出す方法を試行した。

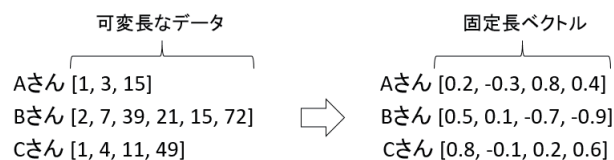


図 4: 固定長ベクトル化

2.4 識別器の構成

最後に、固定長ベクトル化された 120 個の学習データを使用して 120 個の弱識別器を作成する。先にも述べたように、本研究で使用するデータは、全体の約 95% が健康者という不均衡なデータである。ここで、不均衡クラス分類問題において確率論的観点から有効な手法を提案した研究 [Wallace 11] では、不均衡なデータセットに対して UnderSampling と Bagging を組み合わせることで精度向上を見込めることが示されている。そこで我々は、図 3 に示したように UnderSampling したデータに対して Bagging を適用し、多数決による予測を行った。なお、Bagging を施すにあたって、それぞれの弱識別器は決定木をもって構成した。

3. 実験

3.1 データセット

データセットには、健保組合から提供された12年間のデータを用いた。予測対象となる疾病名は表1に示す通りで、重症者の教師ラベルを1、健康者は0とした。データの前処理にあたって、健康者の場合、2017年5月より前のデータがない人を削除し、重症者の場合、対象疾病初出のレセプトから1年前より過去のデータがない人及び対象疾病初出のレセプトから1年前までの間にレセプトが1つもない人を削除する。

3.2 予測設定

退職者・死亡者も含む、全77,665人のうち70%を訓練データとして抽出し、残りの30%をテスト用データとして用いた。形成された学習データらを固定長ベクトル化した後に、120個の枝刈りなしの決定木を作り、多数決によって予測結果を決定した。精度はPrecision, Recall, F1-Scoreの数値によって評価した。固定長ベクトル化にはTF-IDFによる方法とRNNの中間層の出力を取り出す方法の二つを用意し、精度比較を行った。RNNにおいては、Embedding層、Dense層、LSTM層、Dense層、Dense層の、5層のネットワークを構成し、第三層目のLSTM層から256次元の固定長ベクトル化された出力を取り出した。Embedding層への入力時、各学習データは可変長系列なので、最大系列長に合わせて入力直前にゼロパディングする。また、最後のDense層にのみ、活性化関数としてsigmoid関数を適用した。最適化法はRMSpropを、損失関数にはBinary Crossentropyを使用した。

3.3 結果

まず、比較のための従来手法を挙げる。従来手法では、健診のいくつかの項目に閾値を設定し、それらの論理和で重症者か健康者かを区分する。その閾値は、日本人間ドック学会が公式に発表している判定区分表に基づいて決定した^{*1}。設定された閾値は以下の通りである。なお血糖値に関しては、人間ドックと今回の健診の条件が異なるため、HbA1cという項目において過去のデータから参考にした値 θ を基準に判定している。

収縮期血圧: 160以上, 拡張期血圧: 100以上, クレアチニン(男性): 1.30以上, クレアチニン(女性): 1.00以上, eGFR: 44.9以下, 尿酸: 9.0以上, HDLコレステロール: 34以下, LDLコレステロール: 59以下, 180以上, 中性脂肪: 29以下, 500以上, AST: 51以上, ALT: 51以上, γ -GTP: 101以上, 血色素量(男性): 12.0以下, 18.1以上, 血色素量(女性): 11.0以下, 16.1以上, 血小板数: 9.9以下, 40.0以上, 尿蛋白: (2+)以上, HbA1c: θ

この従来手法をベースラインに、TF-IDFによる固定長ベクトル化の方法、RNNの中間層の出力を取り出す方法の精度を比較した。表2に重症者の、表3に健康者の実験結果を示す。

今回試行した固定長ベクトル化の内、どちらの方法も健康者のPrecisionは高いが、重症者の場合RNNによる方法の方が少し高い数値を記録した。またRecallについては重症者、健康者ともにRNNを利用した方が良い結果となった。どちらの方法も健康者においてはPrecision, Recall, F1-Score全ての評価で従来手法のそれに劣る結果となったが、重症者において

は、どちらの方法もRecallが従来手法を上回り、RNNを利用した方法に関してはF1-Scoreが従来手法と同等という結果となった。

表 2: positive(重症者) の精度比較

	Precision	Recall	F1-Score
従来手法	0.11	0.45	0.17
TF-IDF	0.07	0.56	0.12
RNN	0.10	0.65	0.17

表 3: negative(健康者) の精度比較

	Precision	Recall	F1-Score
従来手法	0.98	0.87	0.92
TF-IDF	0.96	0.59	0.73
RNN	0.97	0.64	0.77

表2, 3より、2つの提案手法の内、RNNの中間層を取り出す方法による固定長ベクトル化の方が優位であることが示されたため、この方法で更なる追加実験を行った。我々は、恒川らの研究[恒川19]によって導き出された、識別に有効な健診・問診項目の上位10項目を特徴量に加えた。健診項目からは、HbA1cの値、メタボ判定名、尿糖名、代表判定名、受診年度年齢、心電図判定名の6項目を抽出し、問診項目からは、血圧を下げる薬を飲んでいるか、インスリン注射又は血糖を下げる薬を飲んでいるか、脂質異常症を改善する薬を飲んでいるか、医師から心臓病(狭心症、脳梗塞等)にかかっているといわれたり、治療を受けたことがあるか、という4つの質問を抽出した。健診項目は、HbA1cと年齢が数値で記録され、尿糖名は5段階評価、それ以外は全て6段階評価となっている。問診項目はいずれも、はい又はいいえで回答する形となっている。実験の手順は以下の通りである。まず提案手法で述べた特徴量をRNNを使用して固定長ベクトル化する。次に、先に示した10項目を、one-hot表現として固定長ベクトルに付加する。ただしHbA1cは数値のまま加え、年齢は10代、20代のように年齢層に分割してone-hot表現に変換して加える。こうしてできた特徴ベクトルで決定木を作り、多数決を行った結果が表4である。この結果で、重症者に対するRecallは提案手法の中で最も高い数値を得た。

表 4: 健診・問診項目を追加した結果

	Precision	Recall	F1-Score
positive	0.09	0.75	0.16
negative	0.97	0.56	0.71

4. 考察

表2, 3より、提案手法ではRNNによる方法の方がPrecisionの値が向上している。さらにRecallの観点からも、RNNによる方法の方がTF-IDFを利用する方法と比べて高い数値を出している。これらのことから、時系列性を加味できるRNNの中間層を取り出す方法の方が優位であることが考えられる。しかし、健康者はPrecisionもRecallも従来手法の方が良い結果となった。重症者に関しては、提案手法のRecallが従来手法を超え、F1-ScoreはRNNを使用する方法においては同等

*1 <https://www.ningen-dock.jp/wp/wp-content/uploads/2013/09/Dock-Hantei2018-20181214.pdf>

であったが、Precision はどちらも従来手法に劣る結果となった。このような結果になった原因は、データの形成の仕方にあると考える。普通、患者は自身の体調に異変を感じたり健診結果が悪かったからという理由で受診するが、本研究のように重症者において対象疾病名初出から n カ月前以前のデータを使う、というようにデータを削ると、健康者と同じような内容のデータができるのは必然的なことである。一見健康者のようなデータなのに教師信号は 1 というデータが多数あったため、健康者も重症者も従来手法の Precision に劣る結果になったと考える。しかし逆に、そのような方法でデータの预处理をしているにも関わらず、重症者の Recall は提案手法の方が大幅に高く、RNN を使用する手法では F1-Score が従来手法と同等であることも事実である。このことから、レセプトに自然言語処理を適用する形で提案手法のようにデータを処理することに関して、一定の有効性が認められる。

これらの結果と考察を受け、更なる精度向上を目的とした追加実験では、健診と問診の項目を加えることで重症者に対する Recall の数値を大幅に向上させることに成功した。これは、健診と問診合わせて 57 個ある項目の内から、識別に有効な上位 10 項目を選出したことが精度向上に繋がったと考える。しかし、それでもなお健康者の Recall, F1-Score は従来手法のそれと比較して大幅に低く、重症者に関しても Precision を伸ばすことはできなかった。

ここから精度を上げるためには、各疾病が何に起因して発症するのかを吟味し、それらの情報を特徴量に加えることが必要であると考え。例えば本研究でも対象疾病の一つに加えている“インスリン非依存型糖尿病”は、一般的に遺伝により発症するリスクが高いことが知られている。このように、遺伝性や家族の病歴、喫煙や飲酒、毎日の運動量といった生活習慣に関する情報を特徴量として加えることを、専門家も交えて考えることで精度の向上が見込めると考える。

5. おわりに

本研究ではレセプトデータを用いて複数の疾病名を対象に、今後 1 年以内に発症する可能性のある人の予測を試みた。従来手法と比較して、健康者に対する精度は劣る結果となったが、重症者の Recall には大幅な精度向上が見られた。このことから、レセプトデータに表現学習を用いる考え方は有効であると言える。今後の研究では遺伝性、飲酒や喫煙などの生活習慣に関する情報を特徴量に追加することで精度向上を目指したい。

参考文献

[国土交通省 17] 国土交通省:健康起因事故の発生状況と健康起因事故防止のための取組み
<http://www.mlit.go.jp/jidosha/anzen/03safety/resourse/data/kenkokiinjiko.pdf>, (参照 2019-02-08)

[Ganz 16] Ganz, P., Heidecker, B., Hveem, K., Jonasson, C., Kato, S., Segal, M. R., Sterling, D. G., and Williams, S. A. : Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease., JAMA, Vol. 315, No. 23, pp 2532–2541 (2016)

[Yatsuya 16] Yatsuya, H., Iso, H., Li, Y., Yamagishi, K., Kokubo, Y., Saito, I., Sawada, N., Inoue, M., and Tsugane, S. : Development of a Risk Equation for the Incidence of Coronary Artery Disease and Ischemic

Stroke for Middle-Aged Japanese - Japan Public Health Center-Based Prospective Study., Circulation Journal, Vol.80, No. 60, pp. 1386–1395 (2016)

[Choi 17] Choi, E., Schuetz, A., Stewart, F. W., and Sun, J. : Medical Concept Representation Learning from Electronic Health Records and its Application on Heart Failure Prediction, arXiv:1602.03686, Cornell University (2017)

[Weng 17] Weng, W.-H., Waghlikar, B. K., McCray, T.A., Szolovits, P., and Chueh, C.H. : Medical sub-domain classification of clinical notes using a machine learning-based natural language processing approach, BMC Medical Informatics and Decision Making, Vol. 17, No.1, pp.1–13 (2017)

[Choi 16] Choi, E., Bahadori, T. M., Schuetz, A., Stewart, F.W., Sun, J. : Doctor AI: Predicting Clinical Events via Recurrent Neural Networks, arXiv:1511.05942, Cornell University(2016)

[Chen 17] Chen, M., Hao, Y., Hwang, K., Wang, L.[Lu], and Wang, L.[Lin]. : Disease Prediction by Machine Learning over Big Data from Healthcare Communities, IEEE Access, Vol. 5, pp. 8869–8879 (2017)

[Liu 18] Liu, J., Zhang, Z., and Razavian, N. : Deep EHR: Chronic Disease Prediction Using Medical Notes, arXiv:1808.04928, Cornell University (2018)

[Bo 18] Bo, J., Chao, C., Zhen, L., Shulong, Z., Xiaomeng, Y., and Xiaopeng, W. : Prediction the Risk of Heart Failure With EHR Sequential Data Modeling, IEEE Access, Vol. 6, pp. 9256–9261 (2018)

[Nori 17] Nori, N., Kashima, H., Yamashita, K., Kunisawa, S., and Imanaka, Y. : Learning Implicit Tasks for Patient-Specific Risk Modeling in ICU., Thirty-First AAAI Conference on Artificial Intelligence, pp. 1481–1487 (2017)

[Wallace 11] Wallace, C. B., Small, K., Brodley, E. C., and Trikalinos, A.T. : Class Imbalance, Redux, 2011 IEEE 11th International Conference on Data Mining, DOI:10.1109/ICDM.2011.33, pp. 754–763 (2011)

[恒川 19] 恒川 充, 岡 夏樹, 荒木雅弘, 新谷元司, 吉川昌孝: 健診データを用いた生活習慣病の発症予測, JAMI & JSAI AIM 合同研究会, (2019)