医療情報学会・人工知能学会AIM合同研究会資料
SIG-AIMED-007-10

健診データを用いた生活習慣病の発症予測

Prediction of onset of lifestyle diseases using health checkup data

恒川　充[1*]　岡　夏樹[1]　荒木 雅弘[1]　新谷 元司[2]　吉川 昌孝[3]
Mitsuru Tsunekawa[1]　Natsuki Oka[1]　Masahiro Araki[1]
Motoshi Shintani[2]　Masataka Yoshikawa[3]

[1] 京都工芸繊維大学
[1] Kyoto Institute of Techonology
[2] SG ホールディングスグループ健康保険組合
[2] SG Holdings Group Health Insurance Association
[3] 日本システム技術株式会社
[3] Japan System Techniques Co.,Ltd.

Abstract: This study proposes a method for predicting the onset of lifestyle-related diseases using periodical health checkup data. We carefully examined insurance claims data to identify the onsets of the diseases and used them as correct answers for supervised learning. We adopted the undersampling and bagging approach to address the class imbalance problem. We aimed to predict whether lifestyle-related diseases, other than cancer, will develop within one year. The precision and recall of the proposed method were 0.33 and 0.89, respectively. Compared with a baseline that sets thresholds for each examination item and considers their logical sum, it was found that much higher precision could be obtained while maintaining recall, which is meaningful as it allows for the suppression of the number of targets for health guidance, without increasing the negligence of those that are likely to become severely ill.

1　はじめに

昨今，ネット通販の急速な普及により宅配件数が大幅に増加しており，宅配便運転者の労働環境や健康管理への社会的関心が高まっている．適切な健康指導により，運転者の生活習慣病の発症を減らしたり，運転中の重篤な突発性疾患の発症を防いだりすることができれば，医療費の抑制や交通事故の減少が期待でき，社会的な意義は大きい．そこで本研究は，宅配運送事業者の健康保険組合が持つ健診データから生活習慣病の発症をできるだけ高い精度で予測し，適切な保健指導につなげることを目標とする．

医療データから疾患の発症を予測するために機械学習およびデータマイニング技術を使用している研究は多く存在する． 例えば、日常的な臨床データから心血管リスクを予測するための機械学習技術の優位性を強調した研究 [1] や，健康診断の結果を用いて心筋梗塞または脳梗塞の発生確率を予測した研究 [2] が挙げられる． また， [3] は、定期健康診断データの Lasso ロジスティック回帰を使用して肺炎入院を予測するモデルを提案した。健康な人の定期健康診断データを使用して疾病を予測することを試みている点は本研究と同様である．本研究の目的は、宅配運送事業者の定期健康診断データを使用して、癌以外の生活習慣病が 1 年以内に発症するかどうかを予測することである．

*連絡先：京都工芸繊維大学工芸科学部設計工学域情報工学課程
〒 606-8585 京都府京都市左京区松ケ崎橋上町
E-mail: m-tsune@ii.is.kit.ac.jp

2　データについて

2.1　データの概要

本研究では，SG ホールディングスグループ健康保険組合が持つ従業員のレセプトデータと定期健康診断データを利用した．レセプトとは，患者が受けた保険診療について医療機関が保険者に請求する医療報酬の明細書のことである．例を挙げると，患者の性別や年齢，診療年月といった基本情報をはじめ，診断された病名や診療行為，処方された医薬品などがレセプトデータには含まれている．一方，健診データは，健康診断の結果をまとめてあり，身長，体重，血圧，赤血球数などが記されている．レセプトデータはある人が怪我

もしくは病気にかかり，医療機関で受診した際に作成されるデータであるのに対し，健診データは概ね 1 年に 1 回，定期的に取られるデータである．この二つのデータは，従業員を一意に特定できる匿名のハッシュコードで紐づけされている．本研究では健診データを入力として病気の発症の有無を予測するが，レセプトデータから病気の発症とその時期を抽出し教師データとして用いた．

以下に，データの情報を記載する．健診データは 2006 年〜2018 年のもので，レセプトデータは 1996 年〜2017 年のものを用いた．年齢層は 15 歳〜74 歳，健診データは 156,145 人分の計 961,906 枚，レセプトデータは 108,581 人分の計 1,617,078 枚存在する．

2.2 予測対象とする病名の同定

レセプトデータに含まれている病名コードを見て病名を判断した．病名コードには世界保健機関が作成した疾病及び関連保健問題の国際統計分類コードである ICD-10 を用いた．本研究で予測対象（以降，重症化病名と呼ぶ）とした疾病の ICD-10 コードと病名の対応を表 1 に示している．

表 1: 本研究で予測対象とした疾病の ICD コードと疾病名

ICD-10	疾病名
E10	インスリン依存性糖尿病
E11	インスリン非依存性糖尿病
E14	糖尿病
I20	狭心症
I21,I22	急性心筋梗塞
I42	心筋症
I44〜I49	不整脈，伝導障害
I60,I690	くも膜下出血
I61,I691	脳内出血
I63,I693	脳梗塞

2.3 予測に際して使用した特徴量

予測のための特徴量として利用した健診データの項目を以下に示す．検査結果の数値データだけでなく，生活習慣に関するアンケートの回答結果や，健康診断で測定したデータを用いて医療機関が導き出した六段階の判定結果も健診データの中に含まれている．なお，欠損値の割合が 50 %以上存在した項目である腹囲，心拍数，視力判定，眼底判定，メタボ判定については特徴量から取り除いた．その他にも，健診データには自由記述である医師が記述した所見の内容も含まれているが，自然言語の理解が必要であるため，今回は利用していない．

利用した健診データ項目：
性別／年齢／身長／体重／体脂肪率／収縮期血圧／拡張期血圧／赤血球数／ヘモグロビン／ヘマトクリット／血小板数／ GOT ／ GPT ／γ-GTP ／総コレステロール／ＨＤＬコレステロール／ＬＤＬコレステロール／中性脂肪／尿酸／クレアチニン／ eGFR ／ HbA1c ／血圧を下げる薬を飲んでいるか／インスリン注射をしている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／医師から脳卒中にかかっていると言われたり，治療を受けたりしたことがあるか／医師から慢性の腎不全にかかっていると言われたり，治療を受けたことがあるか／医師から貧血があると言われたことがあるか／現在たばこを習慣的に吸っているか／ 20 歳の時の体重から 10kg 以上増加している／ 1 回 30 分以上の軽く汗をかく運動を週 2 以上の頻度で，1 年以上継続して実施しているか／普段の生活で歩くまたは同程度の活動を 1 日 1 時間以上実施しているか／ほぼ同じ年齢の同性と比較して歩く速度が速いか／この 1 年間で体重の増減が± 3kg 以上あるか／人と比較して食べる速度が早いか／就寝前の 2 時間以内に夕食をとることが週に 3 回以上あるか／夕食後に間食をとることが週に 3 回以上あるか／朝食を抜くことが週に 3 回以上あるか／お酒を飲むか (毎日，時々，飲まないの 3 段階) ／飲酒日の 1 日あたりの飲酒量は清酒に換算してどのくらいか (4 段階) ／睡眠は十分とれているか／運動や食生活などの生活習慣を改善してみようと思うか／生活習慣の改善について保健指導を受ける機会があれば利用するか／尿蛋白判定／尿糖判定／代表判定／身体測定判定／聴力判定／血圧判定／貧血判定／肝機能判定／腎機能判定／尿酸痛風判定／血中糖質判定／糖代謝判定／尿検査判定／診察判定

2.4 データの特徴

データの特徴として，以下の二点が挙げられる．一つ目は，用意されたデータには健康な人のデータが圧倒的に多く存在していることである．例えば，2017 年で，全体のうち重症化病名と診断された人の割合を計算すると，4.5 %でしかなかった．ただ単に全データからランダムに抽出して学習データを作成してしまうとデータ数の多い負例（重症化病名と診断されない人）の特徴が識別結果に強く影響してしまう恐れがあるため，偏りのあるデータをうまく識別できる手法を採用する必要がある．

二つ目は，学習データおよび評価データとして準備する正例データ（今後 1 年以内に重症化病名と診断される健康な人の健診データ）と負例データ（今後 1 年以内には重症化病名と診断されない健康な人の健診デー

タ）の認定が単純ではない点である．本研究では健診の時点で健康な人（重症化病名の診断を受けていない人）に対して，今後 1 年以内に発症するかどうかを予測することを目的とするため，健診時点での病気の有無を正確に判断してデータとする必要がある．健康保険組合が保有するある従業員のレセプトデータに重症化病名が初めて現われた時点が，その人がその病気を発症した時点であるとは限らない．人材の流動が大きい業界では，既に何らかの病気を発症している人が健康保険組合に加入してくる可能性があるからである．使用するデータは一つの企業の健康保険組合のものであるという特性から，データはある人がその健康保険組合に加入している時期の分しかなく，入社前のレセプトデータは確認のしようがない．したがって，保有するレセプトデータで初めて対象病名が現れた時点よりも前の健診データであっても健康な時のものであるとは限らないことになる．この問題への対処法は次節で述べる．

3　データの選定と機械学習手法

3.1　データの選定

今回,「1 年以内に重症化するか否か」を健診データから識別するという 2 クラス分類問題に取り組んだ．このためのデータ選定方法を正例データ，負例データの順に説明する．

そもそも，レセプトデータに記載されている病気の全てを本当の病名の診断と取り扱ってよいとは限らない．なぜなら，ある病気の検査をするためにまだはっきりと病気であると断定できていない状態であってもレセプトに病名を記載する場合があるためである．そこで，いわゆる「疑い病名」と言われるものは病気の診断として取り扱わないようにした．また，それに加えて，病気の治療が実際に行われていれば本当にその病気と診断されたと確定できるので，調剤の情報を確認して，処方されている薬が診断を受けている病気に適応されているものかを見極めた．

まず，前節で述べた問題点に対処するために，重症病名の診断が初めてついたレセプトデータがその人にとって本当に初めての診断らしいか判断する条件を説明する．まず，病気の診断を受けた後の同じ病気での通院間隔を 3 つ計算する．その中で最大の通院間隔より重症病名の診断が初めてついた日と健康保険組合に加入した日との差が大きければ，健康保険組合加入前にその病気を持っていたことはないと見なす．なお，通院間隔のサンプリング数は 3 で十分であろうと判断した．以下に具体的な手順を記す（図 1 も参照）．

1. 重症化病名を持つある人のレセプトデータから，重症化病名が記されている最も古いデータを抽出する．
2. 抽出したデータよりも新しいデータで，同じ病名がついた診療年月が近いデータを 3 つ取り出し，通院間隔を計算する．
3. その人の最も古いレセプトデータ，つまり健康保険組合に加入してから初めてのデータを取り出し，その診療年月と 1 . で抽出したデータの診療年月の差を計算する．
4. 2. で計算した 3 つの値の最大値が 3. で計算した値より小さければ，1. の時点をその人にとって初めて重症化病名だと診断された時点だとみなす．

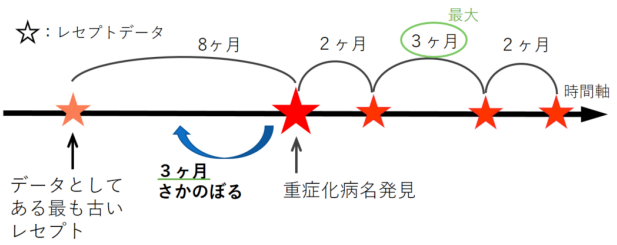

図 1: 病名が初出であるかを判断する条件

次に，正例とする健診データには，初めて重症化病名だと診断されたとみなすデータの診療年月から 1 年前以内の範囲に含まれるデータを選んだ．範囲内に複数のデータが存在する場合は最も古いデータを採用した．また，健診データの変化量に注目し，先ほど取り出した健診データの一つ前の健診データとの差分，二つ前の健診データとの差分を計算して特徴量に加えている（図 2）．病気を発症する際には，健診データ上の何らかの項目に変化があると考えられるので，変化量を明示的に特徴量に加えることで識別精度が向上すると考えた．

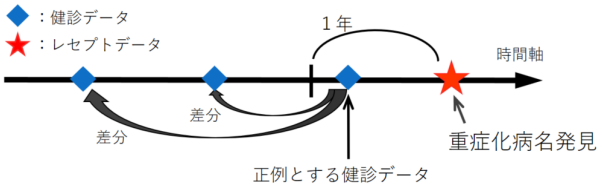

図 2: 正例データの選定

対して，負例データであるが，重症化病名の対象である病名が一度でもついた人を除外し，残った人のデータだけを利用した．また，もし，抽出した健診データから 1 年以上あとにレセプトデータがないとすると，そ

の人は 1 年以内に離職したためデータ上には存在しないだけで，この健診データから 1 年以内に重症化病名と診断されている可能性がある．この可能性を排除するために，抽出した健診データから 1 年以上あとにレセプトデータが存在しない人についてはデータセットから取り除いた．また，正例データと同じように健診データ 3 回分を使って差分を計算して特徴量に追加している（図 3）．

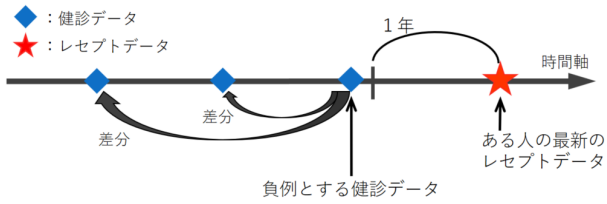

図 3: 負例データの前処理

正例データと負例データに共通して言えることだが，一人の人に対して選定条件に当てはまる複数年分の健診データが存在することがある．そういった場合，データに偏りが生じることを防ぐため，一人につき一つのデータしか利用しないようにしている．正例データは最も古いデータを選択しなければ，重症化病名と診断された後のデータを使ってしまうことになるが，負例データに関してはどの時点のデータを使っても条件から外れることはない．正例データについては，最も古いデータを利用し，負例データは複数個の中から一つだけ任意に選択した．

以上のようにデータの選定を行った結果，正例データが 1255 件，負例データが 37664 件となった．特徴量は全部で 132 であった．欠損値は中央値で埋めているが，50 %以上が欠損値である場合は欠損値処理をしてしまうとデータへの影響が大きいと考え，特徴量から取り除いた．

3.2 用いた機械学習手法

本研究では、不均衡データに対して有効な学習手法として，アンダーサンプリングによりクラス間のデータ数のバランスが取れたデータセットを用意してバギングするという手法 [4] を用いた．手法の概要図を図 4 に示しておく．

これよりバギングとアンダーサンプリングについて，順に詳述していく．バギングはアンサンブル学習の一種で，異なる学習データを複数用意し，それらから複数の識別器を作成し，最後にそれらの結果の多数決をとるという考え方である．多数決をするための複数の識別器のことを弱識別器という．異なるデータセットを複数用意する方法であるが，まず，学習データからランダムにいくつかのデータを取り出し，それをデータセット 1 とする．次に，取り出したデータは元に戻して，また元のデータセットからランダムにデータを取り出し，データセット 2 とする．このようにして弱識別器の数だけデータセットを作成していく．このようなデータセットの作成方法を復元抽出という．復元抽出を行うことにより，様々なデータを持つ異なるデータセットを用意することができる．

今回は，復元抽出の際にデータ件数が少ない正例の数に合わせて負例データをランダムに抽出するというアンダーサンプリングを行った．つまり，複数のデータセットを作った時に，正例データは常に同じものとなり負例データだけが異なったものになるということである．そして，アンサンブル学習の場合，識別器を作成するアルゴリズムが不安定な方が異なる識別器を作り出すことができ，性能が高くなるため，識別器には枝刈りを行わない決定木を使用した．

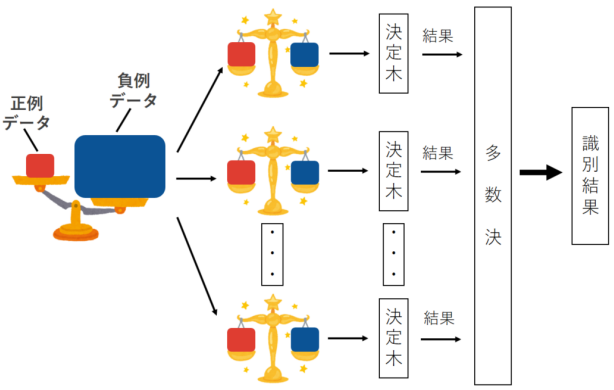

図 4: アンダーサンプリング＋バギング

4 結果と考察

弱識別器数は 500 とした．弱識別器の数を 100 から 500 の範囲で変化させても、recall と precision にはほとんど変化がなかった．しかし，弱識別器の数を 100 未満にすると precision が低下した．弱識別器に利用している決定木は，スケールに影響されないアルゴリズムであるので，データのスケーリング処理は行っていない．評価方法としては，層化 10 分割クロスバリデーションを採用した．まず，全ての対象病名を正例として識別を行ったときの混同行列が表 2 である．正例の precision は 0.33，正例の recall は 0.89 であった．

表 2: 全ての対象病名を正例として識別した場合の混同行列

		予測されたクラス	
		Positive	Negative
実際のクラス	Positive	1118	137
	Negative	2306	35358

ベースライン手法としては，日本人間ドック学会が公表している判定区分表[1]を使用して比較を行った．日本人間ドック学会で用いられている項目の中でも，本研究で利用した健診データと共通の項目であった 13 項目だけを利用して各項目について閾値を設定し，論理和により識別を行った．危険度が高い順に,「要医療，要経過観察，軽度異常」の 3 段階に分類した．

ベースライン手法と提案手法の Precision-Recall 曲線を図 5 に示す．なお，このグラフを描く際は，時間の都合上，データセットのうち 70%を学習に，30%を評価に利用して行った．ベースライン手法は 3 種の閾値による予測結果グラフ上にプロットした．左から，要医療，要経過観察，軽度異常の順である．提案手法では，アンダーサンプリング時の正例と負例の割合を通常は 1:1 でサンプリングするが，左から右に，1:16, 1:8, 1:4, 1:2, 1:1, 1:0.5, 1:0.25 と変化させてデータセットを作ることによって precision と recall を変化させた．識別性能の向上を，採用した機械学習手法を使ったことによる向上と考慮する特徴量を増やしたことによる向上とに分離するため，使用する特徴量をベースライン手法で利用した 13 項目だけに絞って提案手法で識別を行った結果についても Precision-Recall 曲線上に示した．ベースライン手法（特徴量数 13）と提案手法（特徴量数 13）の差が採用した機械学習手法を使ったことによる向上を示し，提案手法（特徴量数 13）と提案手法（特徴量数 132）の差が特徴量を増やしたことによる改善を示す．

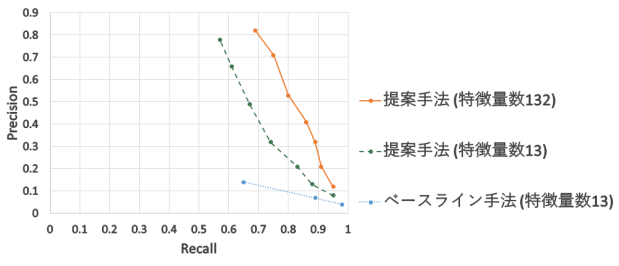

図 5: Precision-Recall 曲線

提案手法のグラフがベースライン手法のグラフよりも上側にあることから，提案手法のほうが優れていることが分かる．つまり，ベースライン手法と同程度の recall の時に，提案手法のほうがより高い precision を得られたと言える．recall を維持しながら precision を向上させることは，疾病を発症する可能性がある人の見落としを増やすことなく保健指導の対象者数を抑えることができ，丁寧な保健指導を実施することを可能にするので，意義のあることである．

識別を行った際にどの特徴量を重視したかを調べた．その結果，特徴量の重要度が高い上位 3 つは，HbA1c，糖代謝判定，インスリン注射または血糖を下げる薬を服用しているか であることが分かった．糖代謝とは，食事として摂取したエネルギーを各臓器が消費して活動し，余分なエネルギーは飢えに備えて蓄え，必要なときに利用するというサイクルのことである．この糖代謝が正常に行われているかを 6 段階で判定したものが糖代謝判定だ．糖代謝が異常をきたすと，糖尿病へと発展していく．また，HbA1c は，糖尿病の判定に用いられる指標の 1 つであり，インスリン注射や血糖を下げる薬を処方するのも糖尿病に関する処置である．このように，糖尿病に関する項目ばかり上位に来ていることから糖尿病は健診データから識別しやすいと考えられる．正例データのうち，糖尿病の数は 73 %にも上るので，糖尿病が識別できていれば全体としても高い精度が出るものと考えられる．

これを確かめるために，糖尿病だけを識別してみた．正例を糖尿病と診断される人とし，負例には糖尿病以外の重症化病名対象者と健康な人のデータを使った．このようにしてデータセットを作成すると，正例データは 921 件，負例 37998 件となった．識別を行うと，正例の precision は 0.32，正例の recall は 0.91 となった．混同行列を表 3 に示す．

表 3: 糖尿病を正例として識別した場合の混同行列

		予測されたクラス	
		Positive	Negative
実際のクラス	Positive	836	85
	Negative	1812	36186

次に糖尿病と比較するために，糖尿病の次に対象病名の中でデータ数の多い狭心症だけを識別した．先ほどと同様に，正例を狭心症と診断される人とし，負例には狭心症以外の重症化病名対象者と健康な人のデータを使った．このようにしてデータセットを作成すると，正例データは 229 件，負例 38690 件となった．識別結果は，正例の precision は 0.04，正例の recall は 0.89 となった．混同行列を表 4 に示す． precision が低下し，狭心症は識別することが難しいことが分かる．

[1]https://www.ningen-dock.jp/wp/wp-content/uploads/2013/09/Dock-Hantei2018-20181214.pdf

表 4: 狭心症を正例として識別した場合の混同行列

		予測されたクラス	
		Positive	Negative
実際のクラス	Positive	204	25
	Negative	5133	33557

5 結言

5.1 まとめ

本研究では，レセプトデータを根拠に学習データを選定する方法と，健診データから生活習慣病の発症を予測する手法を提案した．全ての対象病名を正例として識別すると，precision は 0.33，recall は 0.89 という良好な結果が得られたが，これは糖尿病が識別しやすいものであることに起因していたと考えられる．糖尿病以外の疾病に対しては，recall は高いものの precision が低下することが分かった．

5.2 今後の課題

今回利用できなかったデータとして、胸部 X 線検査や心電図の結果を見て医師が自由記述をしている所見欄がある．この部分に自然言語処理を施すことで，特徴量に追加することができると考える．また，不均衡データに対処する別の方法として，健康な人のデータを正常データとして用いてモデルをフィッティングし，病気が発症するであろうデータを異常として検知するという異常検知の手法の利用も試みたい．また，糖尿病という診断を受ける前にインスリン注射や血糖を下げる薬が処方されることはないはずであるにも関わらず，糖尿病の発症予測に「インスリン注射または血糖を下げる薬を服用しているか」という項目が利用されていることから，正例と負例の選択処理に不十分な部分がある可能性があり，見直す必要がある．糖尿病の場合は HbA1c という分かりやすい指標が特徴量に含まれているので，糖尿病と診断を受けた 2 年前や 3 年前の健診データを利用して HbA1c が高くなることを予測することも試みる計画である．

参考文献

[1] Weng, F. S., Reps, J.,Kai, J., Garibaldi, M. J., and Qureshi, N.: Can Machine-learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?, PLoS One, 12(4), doi:10.1371/journal.pone.0174944 (2017).

[2] 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, 80(60), 1386-1395 (2016).

[3] Uematsu, H., Yamashita, K., Kunisawa, S., Otsubo, T., and Imanaka, Y.:Prediction of Pneumonia Hospitalization in Adults Using Health Checkup Data, PLoS One, 12(6), doi:10.1371/journal.pone.0180159 (2017).

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

[5] 荒木雅弘: フリーソフトではじめる機械学習入門（第 2 版），森北出版 (2014)