一般口演 | 第40回医療情報学連合大会（第21回日本医療情報学会学術大会） | 一般口演
一般口演14
医療データ解析
2020年11月20日(金) 14:00 ～ 15:40 G会場 (イベントホール・特設会場2)

[3-G-3-04] 行列因子分解を使用した個別患者ごとの疾病予測およびリレーショナルデータマイニング

*住谷 有規[1]、中田 和秀[1]、松田 敦義[2]、荒木 賢二[3] （1. 東京工業大学 工学院, 2. 株式会社ログビー, 3. 宮崎大学 医学部附属病院 病院IR部）
*Yuki Sumiya[1], Kazuhide Nakata[1], Atsuyoshi Matsuda[2], Kenji Araki[3] （1. 東京工業大学 工学院, 2. 株式会社ログビー, 3. 宮崎大学 医学部附属病院 病院IR部）
キーワード：Secondary use of EMR, Disease Prediction, Matrix Factorization, Data Mining

病院への電子カルテの導入が進み，蓄積されたビッグデータから意味のある情報を抽出し活用する「二次利用」が注目されている．一方で医師不足の顕在化により，医療の質の低下や医師の負担の増加が懸念されている．そこで電子カルテの二次利用による，1.新たに得られた知見の臨床への応用，2.患者が発症する疾病の予測・予防，が医師不足によって生じる諸問題への解決の一助となると考えられる．

本研究では行列因子分解（ MF）を適用し，以上の課題に取り組む． MFは推薦システムやテキスト解析等に多く使用され，前者ではユーザーに適合するアイテムを提示する手法として，後者では単語の潜在意味解析を行う手法として知られている．本研究では MFを患者-疾病行列（各患者が患った疾病が入力された行列）に適用し，1.患者ごとに各疾病の発症リスクを算出，2.患者および疾病の特徴表現の獲得および解析，3.患者の属性と疾病の関係性の分析，が可能になることを示す． MFの患者-疾病行列への適用に関する報告は少なく，特に上記1~3を同時に目指すのは初の試みとなる．

2008年～2017年の間に宮崎大学医学部附属病院に来院した患者の属性，発症した疾病のデータを用いた数値実験により，当手法の有効性の検証を行った．各患者について発症リスクの高い疾病のランキングを出力し，患者が実際に患った疾病がどのように予測されたか， Top-k Accuracy（予測順位がトップ kに入るデータの割合）等の指標を用いて評価した．また，獲得した特徴表現を用いて，クラスタリングや潜在意味解析等の分析を行った．

MFによる出力は，同患者属性によるランキング手法等よりも高精度であった．さらに獲得した特徴表現の分析を通し，膨大な電子カルテデータから応用可能な情報を抽出できていることを示した．以上より，医療の質の向上や現場の医師の負担軽減が可能になると期待できる．

行列因子分解を使用した個別患者ごとの疾病予測および リレーショナルデータマイニング

住谷 有規[*1]、中田 和秀[*1]、松田 敦義[*2]、荒木 賢二[*3]
*1 東京工業大学 工学院、*2 株式会社ログビー、*3 宮崎大学 医学部附属病院

Patient Disease Prediction and Relational Data Mining using Matrix Factorization

Yuki Sumiya[*1] , Kazuhide Nakata[*1] , Atsuyoshi Matsuda[*2] , Kenji Araki[*3]
*1 Tokyo Institute of Technology, School of Engineering, *2 Logbii, Inc., *3 University of Miyazaki Hospital

Secondary use of electronic medical records will enable us to "analyze the characteristics and relationships of all kinds of medical events" and "predict and prevent diseases that patients may develop", and they will help solve the problems caused by the shortage of physicians. However, not many studies have achieved these simultaneously. In this study, we will address this issue by applying matrix factorization (MF) to patient-disease relationship data. Since there are concerns about applying existing MF methods to this study, we propose a new MF method (named PCMF). We then apply the PCMF to patient-disease matrix and patient-attribute matrix to attempt to perform disease prediction and relational data mining simultaneously. Numerical experiments show that the output of the proposed PCMF is more accurate than other existing methods (its Top-30 Accuracy is 0.4468). Then, the analysis of the acquired feature expressions and the association between disease and patient attributes showed that PCMF can extract useful information from the vast amount of electronic medical record data. This study is expected to improve the quality of medical care and reduce the burden on physicians.

Keywords: Secondary use of EMR, Disease Prediction, Matrix Factorization, Data Mining

1. はじめに

近年、病院への電子カルテの導入が進み、蓄積されたビッグデータから意味のある情報を抽出し活用する「二次利用」が注目されている。また一方で医師不足の顕在化により、医療の質の低下や医師の負担の増加が懸念されている。そのとき，電子カルテデータの二次利用による、

1.医療事象の特徴や関係性の解析

2.患者が発症する疾病の予測・予防

が医師不足によって生じる諸問題への解決の一助となると考えられる。この 2 つの目的について、これまで多種多様な観点から数多くの研究がなされてきた。1.について、電子カルテデータの統計をとるだけでなく、データマイニングにより医療事象の特徴表現を得る手法が提案されている [1)]。この手法の適用により、疾病同士の類似性や疾病と処置の関連性など、臨床に応用可能な情報を得ることができる。2.について、患者が将来に発症する疾病を予測する機械学習手法が数多く提案されている。特に近年では、深層学習を用いてテキストデータも入力に含めて高精度に予測する手法 [2)]、予測と同時にその根拠を提示する手法 [3)] なども開発され、注目を集めている。ただし、1.と 2.を同時に達成する研究の報告は少ない。また、2.について、複数の疾病の発症を同時に予測することは決して容易ではない。発症する疾病を多クラス分類問題として予測する方法が報告されているが [4) 5)]、複雑に関連し合う疾病同士を別のクラスとして排反的に扱うことが最適であるとは考えにくい。

以上の問題を踏まえて、本研究では 1 つの手法で上記 1.2. を同時に達成することを試みたい。これにより、実務上で行う意思決定をさらに明瞭にすることが期待できる。また、疾病の発症予測の際に、強く関連し合う疾病同士を排反的に扱わない手法を取り入れたい。そこで、本研究は推薦システムや画像・テキストの解析など多岐にわたり適用される「行列因子分解」を適用して、以上の課題に取り組む。

本稿は以下のように構成されている。2 節では、本研究で取り扱う問題を具体的に設定する。3 節では、2 節で設定した問題を解くための既存手法とその懸念を説明し、懸念を解決するための新たな手法 PCMF を提案する。4 節では分析結果として、提案手法の有効性の検証を行いつつ、提案手法が臨床に有益な情報を抽出できていることを確認する。最後に 5 節で結論と今後の課題を述べる。

2. 目的

2.1 問題設定

本研究で扱う問題を具体的に設定する前に、「リレーショナルデータ」を定義し導入する。リレーショナルデータとは、ある 1 つの集合と別の 1 つの集合の関係性（リレーション）を示すデータである。この集合として、ユーザー集合や商品集合など、任意の集合が考えられる。また、それらの関係性を示すものとして、例えば、あるユーザーのある商品の購入実績や閲覧履歴などが挙げられる。このとき、2 つの集合がそれぞれ行と列に対応し、各要素をその関係性とする行列がリレーショナルデータとなる。

本研究では、2 つのリレーショナルデータを扱う。患者集合と疾病集合のリレーショナルデータとして「患者-疾病行列」、患者集合と患者属性集合のリレーショナルデータとして「患者-患者属性行列」を定義する。前者の患者-疾病行列を$X \in \{0,1\}^{I \times J}$とすれば、その要素$X_{i,j}$は、患者$i(\in \{1, \dots, I\})$が過去に疾病$j(\in \{1, \dots, J\})$を発症していれば 1、そうでなければ 0 である。後者の患者-患者属性行列を$Y \in \{0,1\}^{I \times K}$とした場合も同様に、要素$Y_{i,k}$は患者$i(\in \{1, \dots, I\})$が属性$k(\in \{1, \dots, K\})$を性質として持つならば 1、そうでなければ 0 である。なお、属性には例えば性別や年齢などを用いることができる。ここで患者-疾病行列において、「患者がこれまで発症していなかったが将来発症する可能性が極めて高い疾病」についても値は 0 で表されていることに注意されたい。得られているすべての患

者・疾病の関連性から、適切な患者と疾病のリレーショナルデータ$\hat{X} \in \mathbb{R}^{I \times J}$を予測として再構築することが本研究の目的である。4 節で紹介する分析では、高血圧症の患者 3,774 人と疾病 500 種から患者-疾病行列を作成し、患者と疾病のすべての関係性（つまり、発症の有無）について、「将来の可能性」として予測を行う。

2.2 本研究の貢献

本研究では行列因子分解（Matrix Factorization、以下 MF）の手法を適用し、この問題に取り組む。MF は推薦システムやテキスト解析等でよく使用され、前者ではユーザーに適合するアイテムを提示する手法として、後者では単語や文書の潜在意味解析を行う手法として知られている。本研究では MF を患者-疾病行列に適用することで、

a.患者ごとに各疾病の発症可能性を予測
b.患者、疾病、患者属性の特徴表現の獲得および解析
c.疾病と患者属性の関係性の分析

の 3 点の実施を試みる。1 節の冒頭で述べた 2 つの目的について、a.は「2.患者が発症する疾病の予測・予防」に、b.および c.は「1.医療事象の特徴や関係性の解析」に対応している。

関連する研究として、患者-疾病行列を MF の一種である NMF（3.1.1 節で紹介する）によって因子分解し、得られた患者の特徴表現をもとにサポートベクトルマシンで分類を行う eDRAM がある [6]。ただしこの報告においては、得られた患者行列と疾病行列を解析することについてあまり言及されていない。また、MF によってリスク要因を抽出していると報告されているが、これは患者の属性を加味されたものではない。本研究や eDRAM のように、 MF の患者-疾病行列への適用に関する報告は極めて少なく、特に上記 a.,b.,c.を同時に目指すのは初の試みとなる。また、本研究では既存の MF を改善した手法 PCMF を新たに提案する。この手法は 3.1.1 節で紹介する NMF の解釈性の高さと、3.1.2 節で紹介する CMF の拡張性の高さを組み合わせた手法である。

2008 年～2017 年の間に宮崎大学医学部附属病院に来院した患者のデータを用いた数値検証において、提案手法 PCMF が他の手法よりも高精度に疾病の発症を予測していることを示した。さらに、PCMF により得られた特徴表現を解析し、臨床に応用可能な情報が抽出できていることを示した。

3. 方法

3.1 節では、2 節で述べた設定に対する既存の解決策として 2 つの手法を紹介し，それらの手法の懸念事項を示す。3.2 節ではそれらの懸念を解決するため、新たな手法 PCMF を提案する。

3.1 既存手法

3.1.1 Non-Negative Matrix Factorization

非負値行列因子分解（Non-Negative Matrix Factorization、以下 NMF）は、非負値のみからなる行列（非負行列）を、2 つの非負行列の積で表現する手法である [7]。すなわち、非負行列 $X \in \mathbb{R}^{I \times J}$が与えられたとき、$X \approx \hat{X}$となるように、

$$\hat{X} = UV^T$$

を満たす 2 つの非負行列 $U \in \mathbb{R}^{I \times R}, V \in \mathbb{R}^{J \times R}$を学習により求める。ただし、$R$は分析者の意思で決定するハイパーパラメータである。NMF はリレーショナルデータから新たな関連性を予測しつつ、各データの特徴表現を抽出できるという特徴を持つ。さらに、この因子分解によって得られる行列の要素は全て非負であるため解釈性が高い。以上の性質により、NMF は推薦システムや画像、自然言語処理など様々な分野で幅広く用いられている。図 1 に NMF のイメージを示す。

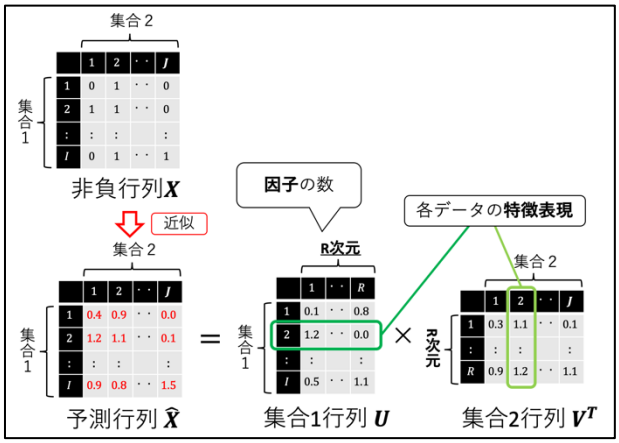

図 1　NMF のイメージ

このNMFを患者-疾病行列に適用する場合、U, Vはそれぞれ患者行列、疾病行列と捉えることができる。すなわち、Uの各行ベクトル$u_1, u_2, \dots, u_I \in \mathbb{R}^R$は各患者のベクトル（特徴表現）を表し、$V$の各行ベクトル$v_1, v_2, \dots, v_J \in \mathbb{R}^R$は各疾病のベクトルを表している。ここで、患者$i(\in \{1, \dots, I\})$と疾病$j(\in \{1, \dots, J\})$の関連性は、

$$\hat{X}_{i,j} = u_i {v_j}^T$$

のように、ベクトルの内積によって特徴づけられる。

ここで、Rは潜在的な因子の数と捉えることができる。例えば$R = 4$とした場合、因子は 4 種類あり、各疾病・各患者がこの 4 つの因子に対してそれぞれ 0 以上の重みをもつこととなる。そして内積の性質から、同じ要素（因子）で大きな重みをもつ患者と疾病同士は、それだけ大きな関連性（発症の可能性）をもつことが推察される。さらに、ある 2 つの疾病ベクトルが類似していれば、その疾病同士はあらゆる患者に対して同時に発症しやすい組み合わせであるということが分かる。これは患者ベクトル同士の組み合わせについても同様に考えられ、性質の「類似性」などと捉えることができる。

ただし、本研究に NMF を適用することに関して懸念されることが 2 点ある。1 つ目は、NMF は 1 つのリレーショナルデータの分析しかできないということである。つまり、1 つの集合と他の集合との関連性を表す別のリレーショナルデータが存在していたとしても、その関連性を加味できない。本研究のように、患者に対して疾病以外の特徴(患者属性）の存在が認められる場合、NMF のこの性質は大きな欠点となる。2 つ目は、本研究で扱う患者-疾病行列の要素は 0 か 1 しか取り得ないため、予測値を内積のみで表現することは柔軟性に欠けてしまうことである。次節で紹介する CMF はこの 2 つの懸念点に関して解決策を提示する手法である。

3.1.2 Collective Matrix Factorization

集合的行列因子分解（Collective Matrix Factorization、以下 CMF）は、ある集合が複数の関係性を持つ場合に、2 つ以上のリレーショナルデータ（行列）を同時に因子分解する手法である [8]。これらの因子分解において、複数の関係性を持つ集合の特徴表現（行列）は共有される。また、各行列の要素が異なる分布を持つ場合にも対応させるため、出力の際に非線形変換を許容している。

本稿においては、CMF を 2 つの非負行列$X \in \mathbb{R}^{I \times J}, Y \in \mathbb{R}^{I \times K}$に対して適用することを考える。ここで$X$と$Y$は、患者-疾

病行列と患者-患者属性行列のように、1 つの集合を共有しているものとする。$X \approx \hat{X}, Y \approx \hat{Y}$ となるように、

$$\hat{X} = f_1(UV^T),\ \hat{Y} = f_2(UZ^T)$$

を満たす 3 つの行列 $U \in \mathbb{R}^{I\times R}, V \in \mathbb{R}^{J\times R}, Z \in \mathbb{R}^{K\times R}$を学習により求める。ここで、$R$は前節同様のハイパーパラメータである。また、$f_1, f_2$ はリンク関数であり、入力と出力の間の非線形的な関係を許容する(本稿では単調非減少な関数を前提とする)。これは行列の各要素に適用する関数であり、入力した行列に対し同じサイズの行列が出力される。図 2 に CMF のイメージを示す。

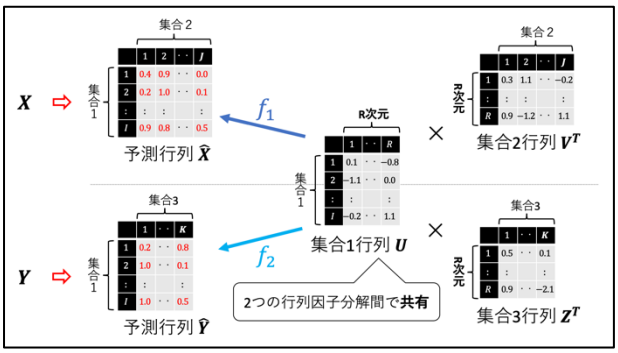

図 2　CMF のイメージ

患者の属性も考慮するため、患者-患者属性行列を同時に扱いたい本研究においては、このように 2 つ以上の行列を同時に因子分解できる性質は望ましいと考えられる。この理由として以下の 2 点が挙げられる。1 つ目は、それぞれの因子分解において患者行列Uは共有されるため、「疾病との関連性」と「患者属性との関連性」が互いに補完し合うことで、より表現力の高い患者の特徴表現を得ることができるようになることである。2 つ目は、疾病と患者属性の関連性を分析ができることである。前節で述べた「類似性」を疾病と患者属性の組み合わせに関しても当てはめることが可能である。一方、元の行列の要素の値が$\{0,1\}$であるため、次節で説明するシグモイド関数を用いれば出力を $[0,1]$に非線形変換できることも、NMF と比較して利点となる。

しかしこの CMF にも、本研究への適用に関して懸念されることが 2 点ある。1 つ目は、不均衡データを扱う場合に生じる問題である。例えば、元の行列の要素の値$\{0,1\}$のうち、1 の数が極端に少ない場合には、0 と予測されやすくなる。本研究において、患者-疾病行列は一般的には 0 が多く 1 が少ない不均衡データとなる。これを解決する有力な手法として、要素ごとに異なる学習率を用いて調整するということが挙げられる。すなわち、少数派のデータに対して学習率を高くする。

2 つ目の懸念点は、得られる行列の要素に正と負の値が同時に出現することである。患者iが疾病jを発症する可能性は、ベクトルu_iとベクトルv_jの内積をとり、単調非減少な変換を行うことによって予測する。この予測に対する解釈・解析を後から行うことは NMF のように容易ではない。例えば、得られた特徴表現について患者iが 3 番目の要素で大きな正の値をとっていたとしても、疾病jが 3 番目の要素で負の値をとれば、それだけ発症の可能性が低くなることを示している。つまり、得られた患者ベクトルの各要素の大きさを、対応する因子によって引き起こされる発症可能性の大きさとして捉えることはできない。このことを踏まえると、行列因子分解によって得られるすべての行列は非負行列であることが、解析には望ましいということが分かる。

以上の NMF、CMF のそれぞれの懸念を同時に解決すべく、次節では新たな手法を提案する。

3.2 提案手法:Positive CMF

3.1 節では、行列因子分解の既存手法として NMF、CMF を紹介した。それぞれに長所だけでなく懸念点が存在し、予測と解釈・解析を同時に行うという本研究の目的の達成のためには、先述の懸念点を克服する手法が必要となる。

そこで本研究では、Positive CMF(以下 PCMF)という手法を新たに開発し、提案する。本研究の問題設定を前提として、PCMF を、患者-疾病行列$X \in \{0,1\}^{I\times J}$と患者-患者属性行列$Y \in \{0,1\}^{I\times K}$に対して適用することを考える。$X \approx \hat{X}, Y \approx \hat{Y}$ となるように、

$$U' = \zeta(U),\ V' = \zeta(V),\ Z' = \zeta(Z)$$
$$\hat{X} = \sigma(U'V'^T - W_X),\ \hat{Y} = \sigma(U'Z'^T - W_Y)$$

を満たす 3 つの行列 $U \in \mathbb{R}^{I\times R}, V \in \mathbb{R}^{J\times R}, Z \in \mathbb{R}^{K\times R}$を学習により求める。ここで、$R$は前節同様のハイパーパラメータ、$W_X \in \mathbb{R}^{I\times J}, W_Y \in \mathbb{R}^{I\times K}$はそれぞれ全ての要素がハイパーパラメータ$w_X, w_Y (\geq 0)$である行列である。また、$\zeta(x) = \log(1 + e^x)$はソフトプラス関数、$\sigma(x) = 1/(1 + e^{-x})$ は(標準)シグモイド関数と呼ばれる活性化関数である。ソフトプラス関数はすべての値xに対し正の値をとる単調増加関数であり、シグモイド関数はすべての値xに対し区間$[0,1]$に変換する単調増加関数である。式中ではこれらの関数に行列を入力しているが、行列の各要素に対して適用し、同じサイズの行列が出力される。

この提案手法の肝所は、各行列にソフトプラス関数を適用することで、すべての要素が正である行列に変換しているという点である。患者行列、疾病行列、患者属性行列をそれぞれU', V', Z'とすれば、NMF で得られる非負行列と同様に解釈を行うことができる(このことはシグモイド関数による変換が単調増加であることより保証されている)。図 3 に PCMF のイメージを示す。

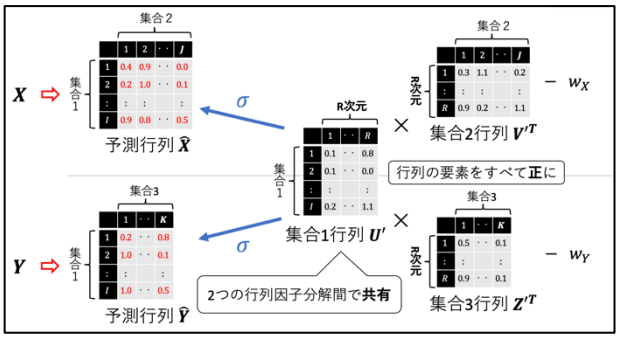

図 3　PCMF のイメージ

この提案手法について、留意すべき事項を 3 点述べる。1 つ目は、W_X, W_Yの意義についてである。$U'V'^T, U'Z'^T$の要素がすべて 0 以上であるため、シグモイド関数による出力を調整するためにこれらを用いる必要がある。これは、入力が 0 以上の場合は出力値が 0.5 以上であるというシグモイド関数の性質に対する措置である。2 つ目は、前節で述べた不均衡データに対する学習である。本研究においては、患者-疾病行列において 1 の値をとる要素について学習率を$\nu (\geq 1)$倍するという方法で対処する(νはハイパーパラメータである)。3 つ目は、解の更新についてである。既存手法の NMF や CMF では解の更新アルゴリズムを解析的に導出する方法が多く提案・採用されているが、本研究における PCMF では深層学習

で使用される誤差逆伝播法 [9]を用いて各行列 U,V,Zの更新を行う。更新のための最適化手法として、Adam [10]など多くの手法を用いることができる。

本研究では、このPCMFを患者-疾病行列$X \in \{0,1\}^{I\times J}$と患者-患者属性行列$Y \in \{0,1\}^{I\times K}$に適用し、新たに患者と疾病のリレーショナルデータ$\hat{X} \in \mathbb{R}^{I\times J}$を予測として再構築することを提案する。適用事例として、患者に対して特に発症の可能性が高い疾病のリストを提示し、医師・患者の双方に発症の予防に努めてもらうことが考えられる。また、得られた患者行列、疾病行列、患者属性行列をもとに、医師が様々な解釈・解析を行うことも想定している。この解析の方法について、次節で具体例を示す。

4. 分析結果

4.1 設定

実電子カルテデータを用いて数値実験を行い、提案手法の有効性の検証、および得られた結果の解釈・解析(リレーショナルデータマイニング)を行う。本節では、数値実験の設定に関して説明する。

4.1.1 データの前処理

本研究では、2008 年～2017 年の間に宮崎大学医学部附属病院に来院した患者のデータを用いて数値実験を行う。

対象とする患者は、1年以上にわたって宮崎大学医学部附属病院で診察を受けており、かつ、これまで対象疾病を 2 つ以上診断されている患者とした。これは、ある程度疾病の情報をもつ患者でないと MF で新たな関係性を予測することが困難であるためである。また、今回は特に「高血圧症」と診断されたことのある患者 3,774 人に対象を絞って実験を行う。高血圧症の患者はあらゆる疾病を引き起こしやすいため、疾病同士の関連性を分析することの重要性が高い。

疾病は ICD-10 に基づき、分類を行う。ICD は集計された死亡や疾病のデータの記録、分析、解釈及び比較を行うため、世界保健機関が作成した分類であり、ICD-10 は ICD の第 10 回目の改訂版である [11]。本実験においては、対象患者内で特に出現頻度の高い 500 種の ICD-10 分類を対象疾病として扱うこととする。ただし、U,V,W,X,Y,Z から始まるコードは除外している。

最後に、患者属性として、「男性」、「女性」、「高齢」(生まれが 1940 年代以前である)、「肥満」(BMI が 25 以上である)、「痩せ」(BMI が 18.5 未満である)の 5 つを用いる。本実験では以上の 5 種類を採用しているが、患者に対してより多くの属性を付与できれば、それだけ高精度な予測や深い解釈が行えるようになる。このようにして、患者-疾病行列$X \in \{0,1\}^{3774\times 500}$、患者-患者属性行列$Y \in \{0,1\}^{3774\times 5}$の 2 種類のリレーショナルデータが得られる。

患者全体のうち検証患者、テスト患者として 15%ずつ割り当て、Xの要素について、その患者らが最後に患った疾病を 0 でマスキングする(以下、この疾病を「マスク疾病」と呼ぶ)。

4.1.2 評価方法

前節の「マスク疾病」の予測値を高く提示したか、という観点で各手法の予測精度を評価する。その指標として、Top-30 Accuracy を用いる。患者ごとに特に予測値が高い疾病を 30 個提示し、その中に「マスク疾病」が含まれている割合を意味する。本実験では、各手法についてパラメータチューニングを行い、検証患者(15%)で最高精度であった学習済みのモデルでテスト患者(15%)を評価し、各手法の比較を行う。

4.1.3 比較手法

以下の手法を適用し、予測精度を比較する。ただし、解釈性を確保するという本研究の目的より、3.1.2 節で紹介した CMF は比較対象から除外する。

- **同性別・年代によるランキング集計**:同じ性別・年代の患者の中で発症した疾病のランキングをとり、それを提示するルールベースの手法。
- **Random Forest** [12]:比較的安定して高いパフォーマンスを発揮することができる教師あり機械学習の手法。また、SHAP value [13]を用いれば、予測に寄与した変数の寄与度を得ることができる。本実験においては、患者の疾病および属性から「マスク疾病」を予測できるように、500 クラスの多クラス分類問題として学習する。特徴量として、目的変数以外の疾病と患者属性の情報を用いる。scikit-learn [14]のモデルを使用。
- **NMF**:3.1.1 節で紹介した手法。MF の性質により、500 種の疾病同士を排反的に扱わない予測が可能。また、得られる特徴表現の解釈も容易。ただし、患者-患者属性行列は加味できない。機械学習ライブラリ scikit-learn のモデルを使用。
- **PCMF**:3.2 節で新たに提案した手法。患者-患者属性行列を用いない場合と用いる場合でそれぞれ学習・予測を行い、比較する。深層学習のフレームワークである TensorFlow [15]で実装を行い、最適化の方法として Adam [10]を用いる。

Random Forest、NMF、PCMF の機械学習モデルにおいて行うパラメータチューニングについて、公平性の観点から探索数を 128 に揃えることとする。

4.2 予測精度の評価

Top-30 Accuracy によって予測精度を比較した結果を表 1 に示す。

表 1　予測精度の比較

手法	Top-30 Accuracy
同性別・年代ランキング	0.3528
Random Forest	0.3883
NMF	0.4131
PCMF(患者-患者属性行列なし)	0.4060
PCMF(患者-患者属性行列あり)	**0.4468**

この表から、提案手法の PCMF(患者-患者属性行列あり)が最高精度であることが分かる。ここで、Random Forest よりも MF の手法である NMF や PCMF が高精度であった理由は、前者は多クラス分類として疾病同士を排反的に扱っており、疾病同士の関連性や共起性を捉えきることができなかったためであると推察される。また、PCMF 同士でも患者-患者属性行列を用いる場合のほうが高精度であり、患者に対して疾病だけでなく属性を加味することの重要性を改めて確認できる。

4.3 リレーショナルデータマイニング

この節では、前節で予測精度 0.4468 を示した学習済み PCMF から得られる患者行列、疾病行列、患者属性行列(およびそれぞれの特徴表現)の解釈・解析を行う。ただし、要素(因子)の数Rは 12 である。

4.3.1 因子の意味解析

はじめに、得られた特徴表現の各因子の分析を行う。特徴表現の 12 個の要素(因子)において特に大きな値をもつ疾病

を抽出し、それぞれの因子の傾向・特徴を分析する。表 2、3 にそれぞれ 3 番目の因子、8 番目の因子を例として提示する。

表 2　3 番目の要素（因子）で大きな値をとった疾病 5 種

ICD-10	疾病名	値
H52.2	乱視	7.335
H40.9	緑内障，詳細不明	5.688
H35.3	黄斑及び後極の変性	5.661
H40.5	その他の眼疾患に続発する緑内障	5.152
H26.9	白内障，詳細不明	5.090

表 3　8 番目の要素（因子）で大きな値をとった疾病 5 種

ICD-10	疾病名	値
C22.0	肝細胞癌	5.315
B18.2	慢性 C 型ウイルス性肝炎	4.880
I85.9	出血を伴わない食道静脈瘤	4.846
K74.6	その他及び詳細不明の肝硬変	4.773
I86.4	胃静脈瘤	4.720

表 2、3 から、3 番目の要素には「眼の疾病」、8 番目の要素には「肝臓の疾病およびそれに関連する疾病」が多く出現していることが分かる。他の 10 の要素に関しても同様の操作を行うことで、対応する因子のもつ意味の解析が可能である。補遺に各要素において値の大きい疾病の一覧を示す。

各因子のもつ意味が分かることで、患者・疾病・患者属性のもつ特徴に関して解析を行うことが可能になる。例えば、ある患者や患者属性の特徴表現について大きな値をとる要素（因子）があれば、その因子に由来する疾病の発症の可能性が高いことが判明する。

4. 3. 2 患者属性の因子の解析

前節で、患者・疾病・患者属性のもつ特徴表現の解析に関して言及した。その具体的事例として、各患者属性と各因子の関連性について、可視化を行う。図 4 に「男性」「女性」、図 5 に患者属性「肥満」「痩せ」の特徴表現における各要素（因子）の値の大きさを示す。

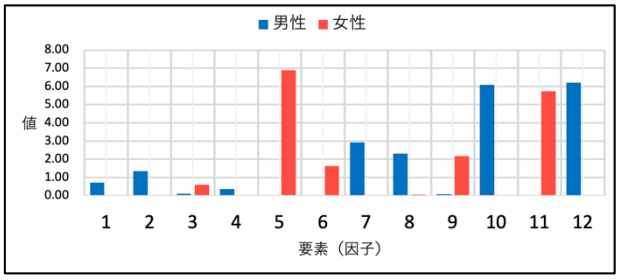

図 4　「男性」「女性」の特徴表現の各要素（因子）の値

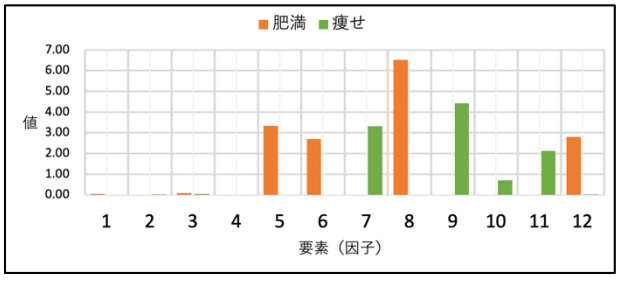

図 5　「肥満」「痩せ」の特徴表現の各要素（因子）の値

特に、「男性」「女性」および「肥満」「痩せ」の関係性はそれぞれ排反的であり、因子の大きさの傾向も異なっていることが窺える。

4. 3. 3 疾病同士の類似性解析

4.3.1 節で各疾病の特徴表現の解析ができることを示したが、同時に疾病同士の関連性を分析することも可能である。すなわち、ある疾病のベクトル（特徴表現）と近い方向性を持つ他の疾病ベクトルがあれば、それらはあらゆる患者に対して同時にリスクとなりやすい疾病同士であると考えられる。これを疾病同士の「類似性」と呼ぶこととする。

ここでは一例として、E11（2型糖尿病）と類似する疾病を抽出する。コードに「E11」を含まないすべての疾病ベクトルに対して、E11 のベクトルと cos 類似度を計算し、特に値の大きい疾病を表 4 に示した。ただし、行ベクトル$a, b \in \mathbb{R}^d$の cos 類似度は以下の式で計算されるものであり、1 に近いほど類似していると捉える。

$$\mathrm{Similarity}(a,b) = (ab^T)/(\sqrt{aa^T}\cdot\sqrt{bb^T})$$

また、表には参考として、各疾病を発症した患者の中で E11 を発症している患者の割合（「E11 発症率」）も示している。ただし、全体での E11 発症率は 0.117 であった。

表 4　E11（2型糖尿病）と特に類似している疾病 6 種

ICD-10	疾病名	E11 発症率	類似度
E14	詳細不明の糖尿病	0.208	0.970
E78.0	純型高コレステロール血症	0.286	0.893
I67.2	脳動脈のアテローム硬化	0.364	0.887
I70.9	全身性及び詳細不明のアテローム硬化	0.163	0.836
I65.2	頚動脈の閉塞及び狭窄	0.216	0.832
K21.0	食道炎を伴う胃食道逆流症	0.135	0.829

表 4 から、E11 と同じく糖尿病である E14 や、糖尿病に起因する疾病が抽出できていることが窺える。

4. 3. 4 疾病と患者属性の類似性解析

前節で疾病同士の類似性が分析できることを示したが、疾病と患者属性の間でも同様に類似性を分析することが可能である。前節同様、ある患者属性ベクトルとすべての疾病ベクトルの間で cos 類似度を計算し、値の大きな疾病を抽出する。ここでは一例として、患者属性「男性」と「高齢」について類似度の高い疾病を、それぞれ表 5、6 に示した。表には参考として、各疾病を発症した患者の中での「男性割合」「高齢割合」も示している。ただし、全体での男性割合は 0.567、高齢割合は 0.621 であった。

表 5　「男性」と特に類似している疾病 5 種

ICD-10	疾病名	男性割合	類似度
K70.9	アルコール性肝疾患，詳細不明	0.944	0.807
I25.2	陳旧性心筋梗塞	0.803	0.807
N40	前立腺肥大（症）	0.997	0.800
I71.4	腹部大動脈瘤，破裂の記載がないもの	0.817	0.789
J43.9	肺気腫，詳細不明	0.914	0.784

表 6 「高齢」と特に類似している疾病 5 種

ICD-10	疾病名	男性割合	類似度
K40.9	一側性又は患側不明の鼠径ヘルニア，閉塞及び壊疽を伴わないもの	0.885	0.865
B90.9	呼吸器及び詳細不明の結核の続発・後遺症	1.000	0.861
H81.0	メニエール病	0.813	0.812
I72.3	腸骨動脈瘤及び解離	0.909	0.799
D37.2	口腔及び消化器の性状不詳又は不明の新生物，小腸	0.733	0.789

他の患者属性である「女性」「肥満」「痩せ」に関しても同様の分析が可能であり、また、患者属性に新たな要素を加えれば、分析の幅をさらに広げることができる。

以上4節全体で紹介した分析は「高血圧症」の患者に対象を絞った上での結果であったが、これを他の特徴を持つ患者、或いは全患者に対して適用することも可能である。その分析によって、また新たな知見が得られることが期待できる。

5. 結論

医師不足によって生じる諸問題の解決のため、「1.医療事象の特徴や関係性の解析」「2.患者が発症する疾病の予測・予防」を同時に達成するべく、本研究を行った。本研究では行列因子分解（MF）に基づいた手法によって、「a.患者ごとに各疾病の発症リスクを算出」「b.患者、疾病、患者属性の特徴表現の獲得および解析」「c.疾病と患者属性の関係性の分析」の実施を実施した。また、既存のMFの懸念を解消するため、PCMF という手法を新たに提案した。そして実電子カルテデータを用いた分析により、提案手法 PCMF が他の手法よりも高精度に疾病の発症を予測していることを示した。さらに得られた特徴表現を解析し、臨床に有益な情報が抽出できていることも示した。本稿では紹介しきれなかったが、「疾病の2次元マッピング」「患者のクラスタリング」も可能であり、これらも分析者に新たな情報を提供することが期待できる。

今後の課題として 3 点挙げられる。1 つ目は、時系列性や因果性を考慮することである。本研究の MF では疾病同士、あるいは疾病-患者属性の間の「相関性」を捉えることはできても、「時系列性」「因果性」を明示的に扱うことはできない。これらを考慮することができれば、より深い洞察を得ることが期待できる。2 つ目は、疾病のカテゴリを考慮することである。ICD-10 などの階層構造の情報を疾病とのリレーショナルデータとして与え、これを含めて PCMF を適用すれば、より精密な疾病行列が得られると考えられる。3 つ目は、実際に医療実務の中に組み込むために、システムの設計・開発・運用を行うことである。予測を行いつつ自動的にアラートメッセージを送信するシステムや、解釈・解析をインタラクティブに行えるツールなど、医師により良い支援を行うための工夫を施したい。

参考文献

1) Choi, E., Bahadori, M. T., Searles, E., et al. Multi-layer representation learning for medical concepts. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016 : 1495-1504.
2) Alvin Rajkomar, Eyal Oren, Jeffrey Dean, et al. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 2018 ; 1.1 : 18.
3) Choi, E., Bahadori, M. T., Sun, J., et al. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In Advances in Neural Information Processing Systems, 2016 : 3504-3512.
4) Choi, E., Bahadori, M. T., Song, L., at al. GRAM: graph-based attention model for healthcare representation learning. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017 : 787-795.
5) Lipton, Z. C., Kale, D. C., Elkan, C., and Wetzel, R. Learning to Diagnose with LSTM Recurrent Neural Networks. arXiv preprint arXiv:1511.03677, 2015.
6) Chin, Chu-Yu, Sun-Yuan Hsieh, and Vincent S. Tseng. eDRAM: Effective early disease risk assessment with matrix factorization on a large-scale medical database: A case study on rheumatoid arthritis. PLoS One, 2018 ; 13 : e0207579.
7) Lee, Daniel D., and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 1999 ; 401.6755 : 788-791.
8) Singh, Ajit P., and Geoffrey J. Gordon. Relational learning via collective matrix factorization. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008 : 650-658.
9) Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. Nature, 1986 ; 323.6088 : 533-536.
10) Kingma, Diederik P., and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
11) 厚生労働省. 疾病、傷害及び死因の統計分類, 2020(更新). https://www.mhlw.go.jp/toukei/sippei/index.html (cited 2020-Aug-25)].
12) Liaw, Andy, and Matthew Wiener. Classification and regression by randomForest. R news, 2002 ; 2.3 : 18-22.
13) Lundberg, Scott M., and Su-In Lee. A unified approach to interpreting model predictions. Advances in neural information processing systems, 2017 : 4765-4774.
14) scikit-learn. machine learning in Python — scikit-learn 0.23.2 documentation, 2020(更新). https://scikit-learn.org/stable/ (cited 2020-Aug-25)].
15) TensorFlow. TensorFlow, 2020(更新). https://www.tensorflow.org/ (cited 2020-Aug-25)].

補遺

4.3.1 節に掲載しきれなかった、各要素において大きな値をもつ疾病（上から 5 種ずつ）を以下の表に示す。

要素 1		要素5		要素9	
N18	慢性腎臓病	E04.9	非中毒性甲状腺腫，詳細不明	N03.9	慢性腎炎症候群，詳細不明
N19	詳細不明の腎不全	E78.5	高脂血症，詳細不明	D80.1	非家族性低ガンマグロブリン血症
N18.9	慢性腎臓病，詳細不明	C73	甲状腺の悪性新生物	J17.3	寄生虫症における肺炎
D63.8	他に分類されるその他の慢性疾患における貧血	D25.9	子宮平滑筋腫，部位不明	B59	ニューモシスチス症
N08.3	糖尿病における糸球体障害	D44.1	副腎	N02.8	反復性及び持続性血尿，その他
要素2		**要素6**		**要素10**	
C34.9	気管支又は肺，部位不明	G20	パーキンソン病	N40	前立腺肥大
C34.1	上葉，気管支又は肺	R52.1	慢性難治性疼痛	C61	前立腺の悪性新生物
C34.3	下葉，気管支又は肺	F32.9	うつ病エピソード，詳細不明	C67.9	膀胱，部位不明
E13.9	その他の明示された糖尿病，合併症を伴わないもの	G98	神経系のその他の障害，他に分類されないもの	C16.9	胃，部位不明
G70.0	重症筋無力症	M48.06	腰部脊柱管狭窄症	D37.1	胃
要素3		**要素7**		**要素11**	
H52.2	乱視	J69.0	食物及び吐物による肺臓炎	I34.0	僧帽弁閉鎖不全
H40.9	緑内障，詳細不明	R09.0	窒息	I35.1	大動脈弁閉鎖不全
H35.3	黄斑及び後極の変性	K80.5	胆管炎及び胆嚢炎を伴わない胆管結石	I50.0	うっ血性心不全
H40.5	その他の眼疾患に続発する緑内障	K83.0	胆管炎	I07.1	三尖弁閉鎖不全
H26.9	白内障，詳細不明	K91.8	消化器系のその他の処置後障害，他に分類されないもの	I35.0	大動脈弁狭窄
要素4		**要素8**		**要素12**	
C79.5	骨及び骨髄の続発性悪性新生物	C22.0	肝細胞癌	I47.2	心室頻拍
D70	無顆粒球症	B18.2	慢性C型ウイルス性肝炎	I50.0	うっ血性心不全
C78.0	肺の続発性悪性新生物	I85.9	出血を伴わない食道静脈瘤	I50.9	心不全，詳細不明
R52.2	その他の慢性疼痛	K74.6	その他及び詳細不明の肝硬変	I42.0	拡張型心筋症
K12.1	その他の型の口内炎	I86.4	胃静脈瘤	I48	心房細動及び粗動