

電子診療報酬データの多角的解析

衛藤 亮太[†] 松原 靖子^{††} 山下 和人^{†††} 國澤 進^{†††}

今中 雄一^{†††} 櫻井 保志^{††}

[†] 熊本大学 大学院自然科学研究科 情報電気電子工学専攻

^{††} 熊本大学 大学院先端科学研究部

^{†††} 京都大学 大学院医学研究科 医療経済学分野

E-mail: [†]161d8620@st.kumamoto-u.ac.jp, ^{††}{yasuko,yasushi}@cs.kumamoto-u.ac.jp,

^{†††}{yamashita.kazuto.4m,kunisawa.susumu.2v,imanaka.yuichi.5m}@kyoto-u.ac.jp

あらまし 本論文では、電子診療報酬データを多角的に解析するための深層学習を用いた医療情報分類モデルについて述べる。提案手法は、利用者が受診した医療機関や診断された疾患および手術方法などの多次元情報を含む医療情報が与えられたときに、その医療情報から分類に有用な特徴を発見し、利用者の予後を表す転帰情報の分類を行う。実データを用いた実験では、提案手法が多次元医療情報の中から分類に有用な情報を特徴間の共起影響を加味した上で特定することを確認し、既存の分類モデルとの比較を行い提案手法の精度が向上をしていることを示した。

キーワード 医療情報, 深層学習, 分類器, 属性選択

1. ま え が き

近年、先進国を中心として高齢者人口の増加や出生率の低下により少子高齢化の進行、ならびに国民医療費の増大といった数多くの問題が社会保障制度に大きく影響を与えている。特に我が国は、先進国の中で高齢化率が最も高く、持続可能な社会保障制度を構築するための対策が急務とされている。これらの対策の一つとして挙げられているのが、昨今の情報化社会の発展に伴い生じた電子カルテの医療情報やCT, MRIなどの診断機器から取得した様々な大規模医療情報の活用である。より具体的には、電子カルテ等から取得した医療情報を利用し患者の容体予測や医用画像解析による診断補助が期待されている。また医療情報を用いた研究の中でも、2012年にILSVRC^(注1)で発表された従来手法を大きく上回る画像認識精度を得た深層学習を利用した研究事例が特に注目されている。深層学習は機械学習アルゴリズムの一つであり、多層ニューラルネットワークを学習することでデータの高次表現を獲得することが可能なことで知られており、画像認識や音声認識の分野で数多くの研究事例が報告されている。しかし、医療情報に対し深層学習を適用した研究事例はCTやMRI画像などの医用画像を解析対象とするものが非常に多い。これは医用画像を用いる場合、画像認識能力に長けた畳み込みニューラルネットワークの適用が所望の結果を得ることが広く認知されていることに起因している。

これらの情報技術は医療に対して大きな役割を果たしているが、一方で医療機関や社会全体の医療健康増進のための有力な医療情報解析手法は未だ確立されていないのが実情である。今後は、情報技術を医療の質の向上や医療資源配分の適正化等に繋げるため、複雑な医療情報を多角的に解析可能な技術が社会

に重要な役割を果たすと考えられる。

そこで本論文では、京都大学医療経済学分野により実施されているQIPプロジェクトより提供される多次元医療情報を含む、個々の情報が匿名化された電子診療報酬データを多角的に解析することを目的とする。より具体的には、病名や手術などの多様な情報を扱い患者の予後を示す転帰の分類を行うため深層学習を用いた分類モデルを提案する。また医療情報の解析に伴い分類を行うモデルへの入力情報の選定は非常に重要な課題である。そのため本論文では、分類モデルの提案に加えて電子診療報酬データから分類に有用な特徴の選定を行う。

本論文の貢献 本論文では、診断された傷病名や手術名等の23種の属性情報で構成される電子診療報酬データ $\mathbf{X} = \{\mathbf{x}_{i,j}\}$ が与えられたとき深層学習に基づく電子診療報酬データのための分類器を構築し、分類に有効な属性を選定する多角的解析手法を提案する。提案手法は次の特徴を持つ。

- 提案手法は電子診療報酬データから医療情報を分類するために有効な属性を医療情報間の共起影響を加味した上で選定する。
- 既存の分類器と比較し高精度な医療情報の分類を実現する。

2. 背 景

本章では本論文の技術的背景について主に述べる。本章の構成として、まず2.1節にて医療情報分類に使用する深層学習について述べ、2.2節にて本論文で行う属性選択の基礎となる特徴選択法について述べる。

2.1 深層学習

深層学習は多層のニューラルネットワークを用いた機械学習

(注1) : <http://www.image-net.org/challenges/LSVRC/>

のアルゴリズムである。深層学習は幾つかの方法論の集まりであり、適用する問題によって方法論を使い分けるのが一般的である。画像認識の分野には視野脳の情報伝達システムをニューラルネットに適用した CNN が用いられる。次元削減や特徴抽出問題には深層自己符号器 (Deep Auto Encoder:DAE) 等が適用される。また音声や言語、動画像などの系列データ処理には再帰型ニューラルネットワーク (Recurrent Neural Networks:RNN) が活用される。本論文では RNN を利用したモデルを構成する。RNN は有向閉路を持つネットワークであり隠れ層に閉路を設置することで、過去の入力情報を現行の出力に反映できる。そのため音声や言語といった時系列データに適用されることが多い。また RNN は有向閉路層を時系列順に展開することで、順伝搬型ネットワークとして扱うことができる。

RNN は有向閉路を持ち、各層間で異なる重みパラメータを保持する。次にある変数 t での RNN の出力を示す順伝搬計算を (1) 式に示す。

$$\mathbf{z}^t = \mathbf{f}(\mathbf{W}^{In-Hi} \mathbf{x}^t + \mathbf{W} \mathbf{z}^{t-1}) \quad (1)$$

(1) 式において $\mathbf{z}$ は出力、 $\mathbf{x}^t$ は t における RNN への入力であり $\mathbf{W}^{In-Hi}$ 、 $\mathbf{W}$ は入力-隠れ層間の重み、および有向閉路の重みを示す。(1) 式から、RNN は過去全ての入力情報を反映した結果が得られると考えられるが、通常の RNN は順伝搬型ネットワークの勾配消失問題によって反映可能な過去情報数に限りがある。

2.2 特徴選択

分類器を作成する際、どのような情報を分類器の学習に用いるかという問題はクラス間で違いを示す医療情報を見つけることに相当し分類器の性能に大きく影響する。特徴選択法は入力情報が保有する性質を利用し、ピアソンの相関係数や相互情報量を指標として用いる Filter 法、そして選択する特徴部分集合を変化させながら各特徴の評価を行う Wrapper 法に大別される。一般的に Wrapper 法は、実際の分類器により特徴の評価を行うため各特徴の評価である識別率は高くなるが、多量の分類器を構成するため計算量が多い。一方で Filter 法は計算量は少ないが識別率は低くなるとされている。特徴選択を行う上では各手法とタスクの性質を鑑みた上で計算量と識別率のトレードオフを選択する必要がある。また電子診療報酬データの多次元医療情報の分類においては、分類対象である転帰等の種別の多様性に影響を受けない特徴に加えて、複数の特徴による共起影響を考慮する必要がある。複数の特徴の共起影響を考慮した Filter 法に基づく特徴選択手法として MCFS (Multi Cluster Feature Selection) [1] がある。MCFS はデータセットから分類精度に起因する特徴を選択することができる。しかし、本論文で用いる電子診療報酬データは名義変数型の情報が多数存在すると共に、多数の傷病名から構成される病名のような複数の特徴から構成される属性情報の選択を行うことができない。次に本論文の関連研究について述べる。

3. 関連研究

本論文の関連研究として医療情報を情報技術によって解析した研究事例および深層学習、特徴選択の適用研究について主に述べる。

3.1 医療情報解析

医療情報に対して情報処理を適用した研究は X 線や MRI などの医用画像に加えて、心電図や脳波といった生体信号処理など多岐に渡って行われている。また近年、本論文で扱う電子診療報酬データに類する電子カルテ (Electical Health Record:EHR) の普及率向上に伴い解析対象として扱われる事例が増加しつつある。EHR に関する取り組みは、医療機関連携ネットワークを利用して個々の医療情報の共有を目指すものであり、欧米、そして、わが国でも医療構造改革の目標を達成するための有力な手段として位置付けられている。EHR を解析対象とした研究事例として、文献 [10] らは ICD10 構造を利用したマルチタスク学習に基づいた ICU 利用者の死亡リスク予測を行なっている。特に本研究と関連が深い研究事例として、EHR の医療情報に対してクラス分類を実施した文献 [11] [12] では分類器としてランダムフォレスト (Random Forests) やサポートベクターマシン (Support Vector Machine:SVM) を利用している。特に SVM はパターン認識能力に優れたモデルであり、One against reset 法を用いることによって多クラス分類を実現できる。またカーネル関数を用いて定義される非線形変換によって非線形分類器に拡張できる高い汎化性能を持つ。上記の研究事例での情報技術は医療に対して大きな役割を果たしているが、EHR の疎密性の課題から解析対象とする傷病を限定するなどの処置をとっており、大規模医療情報の一般的な解析手法は未だ確立されていないのが現状である。本論文では対象とする傷病を限定せずに電子診療報酬データから医療情報を分類するための手法を提案する。続いて本論文で利用した深層学習に関する関連研究について述べる。

3.2 深層学習

深層学習は多層なニューラルネットワークであり、層間の結合状態によって順伝搬型ニューラルネットワーク (Feedforward Neural Network)、再帰型ニューラルネットワーク (Recurrent Neural Network:RNN)、無向型モデル (Undirected models) に大別される。本論文では順伝搬型および再帰型のネットワークを利用する。順伝搬型ニューラルネットワークの代表的なモデルとしては、MLP や画像認識に長けた CNN、自然言語処理の手法として用いられる Word2Vec [9] が挙げられる。特に Word2Vec は単語の分散表現を獲得するモデルであり、OneHot 符号化や Binary 符号化などの 1-of-K 表現の符号化法と比較して密な単語表現情報の算出や、単語の分布表現から意味的に関連の高い単語グループのクラスタリングが可能といった利点がある。本論文では電子診療報酬データを符号化する手法として Word2Vec を適用した。また RNN は音声情報などの時系列情報に対して適用事例が多く報告されている深層学習モデルであり、過去情報を加味したモデルを学習することが可能である。RNN を適用した研究事例としては EC サイト等のセッション

データを活用した推薦システム [4] やスケルトンデータの共起影響を加味した行動認識に関する研究事例 [5] など多様な分野への適用、モデルの改良が活発に行われている。本論文においては、電子診療報酬データのための分類器構築の際に RNN を利用した。続いて本論文で行う属性選択に関する関連研究として特徴選択法について述べる。

3.3 特徴選択

本論文では電子診療報酬データから分類に有用な特徴を選択する際に Filter 法, Wrapper 法を適用し、かつ複数の属性の共起影響を考慮した選択を Filter 法によって実現する。そのため Filter 法の関連研究について述べる。特徴選択として Filter 法を適用した事例は数多くあり、局所多様体構造を保つように各特徴のスコアを算出する Laplacian Score [2] や、Spectral Clustering の次元縮約表現に基づき各特徴のスコアを Lasso によって算出した MCFS [1] がある。これらの先行研究はある目的変数の分類やクラスタリングなどのタスクにおいて非常に有効であることが実証されているが、複数特徴から構成される属性の選択を行うことができない。本研究では電子診療報酬データの分類に有効な属性の選択を行う。続いて本論文で提案する手法について具体的に述べる。

4. 提案手法

本章では電子診療報酬データを対象とした提案手法について述べる。提案手法の処理概要を図 1 に示す。まず初めに本論文で扱う電子診療報酬データについて述べる。

4.1 電子診療報酬データ

本論文では、京都大学医療経済学分野により実施されている QIP プロジェクトより提供される匿名化された医療情報データを用いる。QIP では全国の参加病院より匿名化された DPC データの提供を受けており、本論文ではさらに病院などの特定もできない形での実施を行う。DPC データは DPC/PDPS と呼ばれる診療報酬の包括支払い制度に参加する病院で作成が必須となっている全国共通フォーマットデータである。このデータには行った診療行為データのほか、病名や術式、あるいは入退院時の状況が含まれる。これらのデータは文字列表記のほか、傷病は国際標準の ICD10 コード、手術は診療報酬で規定された K コードと呼ばれるものなどによりコーディングもされている。そこで ICD10 並びに K コード等によって記録されている属性を代表して主傷病名 (dmain), 手術名 (opek), 退院時転帰 (outcome) のそれぞれについて述べる。また本論文ではこれらの項目を電子診療報酬データの属性と呼び 23 種の属性をデータとして利用した。

(1) 主傷病名

主傷病名の属性名を dmain とする。dmain は患者が診断された主な疾病の名称が ICD10 コードによって記載されている。ICD(International Statistical Classification of Diseases and Related Health Problems) は、死因や疾病の統計基準として世界保健機構によって作成された国際統計分類の指標である。

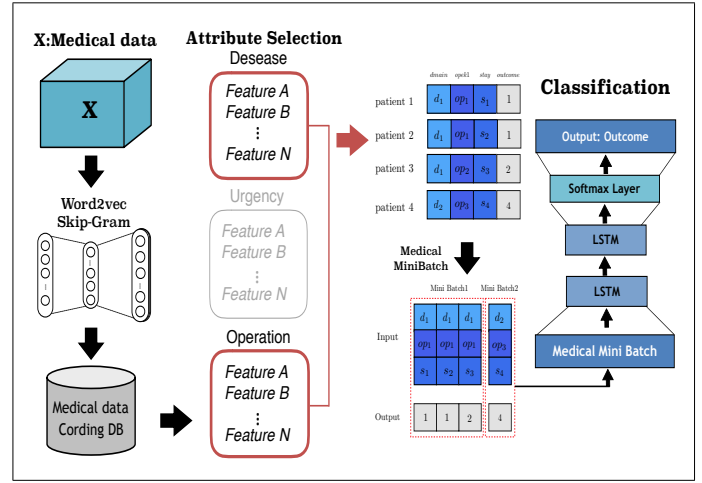


図 1 提案モデルの処理概要

表 1 転帰区分

転帰	区分
1	傷病が治癒したと判断される場合
2	傷病が軽快したと判断される場合
3	傷病 (白血病等) が寛解したと判断される場合
4	傷病が不変と判断させる場合
5	傷病が悪化したと判断される場合
6	傷病による死亡
7	傷病以外による死亡

ICD10 コードは、アルファベットと数字で構成され、例えば”I63.3”という ICD10 コードは脳動脈の血栓症による脳梗塞を示す。

(2) 退院時転帰

退院時転帰の属性名を outcome とする。outcome は患者に対して治療を行った後の病状の変化を数値で示す。転帰の状態は、退院時に主治医の判断に基づいた患者の状態が記載されている。表 1 に各数値の概要を示す。

(3) 手術情報 (opek)

手術情報の属性名を opek とする。opek は患者に対して行った手術情報を示す。手術情報は診療報酬点数表の診療行為の中から、手術および処置の領域に割り当てられた K コードによって符号化され、電子診療報酬データには患者毎に最大 5 件の手術情報が存在する。そこで、それぞれの属性名の接尾語として数値を付け加える。

4.2 医療情報符号化

分類器への入力となる医療情報は ICD10 コード及び K コード等によって記載されている多値論理関数であり直接的に扱うことは難しい。多値論理関数を扱う場合、一般的に複数の 2 値変数を組み合わせた表現技法である 2 進符号化 (binary coding) またはワンホット符号化 (One-hot coding) が用いられる。しかし、両手法では次元数の増大に加えて、符号化後の特徴に対して演算を行うことが難しく属性選択を行うための類似度

行列を作ることができない．そこで本研究では電子診療報酬データの符号化方法として深層学習モデルである Word2Vec Skip-Gram^(注2)を用いる．Skip-Gram は入力層、隠れ層、出力層の三層から構成され OneHot 符号化した入力情報に対して任意の周辺単語を出力層に設定し学習を行う．電子診療報酬データは各カラム毎に利用者の属性が格納されている．そこで Skip-Gram の学習の際には、各利用者毎の医療情報を対象とし、他の利用者の情報を加味せずに処理することで、利用者毎の医療情報を加味した符号化情報を獲得する．また Skip-Gram によって電子診療報酬データの多次元情報を任意次元のベクトル空間に射影することで、各医療情報の類似度の算出が可能になる．例えば、ICD10 コードにより C220 と表現される肝臓に対して高い類似度を持つ要素として C 型肝炎が推定される．上記の結果に対して、肝臓発症の要因の一つとして C 型肝炎ウイルスの感染が挙げられることから先の事例に対して妥当性が伺える．また本論文では電子診療報酬データの各情報を 200 次元のベクトルとして表現した．

4.3 属性選択法

本論文では、電子診療報酬データの分類に有用な特徴を選択する．そこで属性選択法として広く用いられている Wrapper に基づく手法である Top-K を適用する．また電子診療報酬データの病名や手術方法等の医療情報が単体で分類精度に影響するとは考えにくい．そのため電子診療報酬データの各属性間で生じる分類結果への共起影響を考慮した属性選択法である MGAS(Multi Group Attribute Selection) を提案し比較実験を行う．次にそれぞれの手法について具体的に述べる．

Top-K Wrapper 法に基づく属性選択アルゴリズムである Top-K について述べる．Top-K では電子診療報酬データに含まれている全入力属性をそれぞれ提案モデルの分類器への入力とし分類精度が高い属性を順に K 個選択し分類を行う．

MGAS 次に Filter 法に基づく属性選択法である MGAS について述べる．上記は Filter 法に基づくアルゴリズムであり、電子診療報酬データの各属性に対する医療情報スコアを算出し分類器への入力を決定する．次に MGAS の具体的な処理について述べる．

(a) 次元縮約表現

まず MGAS では電子診療報酬データの次元縮約表現を獲得するため、医療情報の Normalized Cut 最小化問題(以下 NCut)を考える．NCut はグラフラプリアン行列およびレイリー商を用いることで一般化固有値問題に帰着する．そこで電子診療報酬データのグラフラプリアン行列 ($\mathbf{L}$) を定義するため類似度行列 $\mathbf{W}_{i,j} = \mathbf{x}_i^T \mathbf{x}_j$ を生成する．類似度行列 (Affinity matrix) は実対称行列であり、 x_i, x_j は文献 [1] に基づきドット積を利用する．また類似度行列から次数行列 $\mathbf{D}_i = \sum_j \mathbf{W}_{i,j}$ を定義する．

次数行列 (Degree matrix) は類似度行列の対角成分の総和であり、類似度行列との差によってグラフラプリアン行列 $\mathbf{L}$ が定義される．以上より電子診療報酬データ $\mathbf{X}_{i,j}$ の多様体学習に基づく次元削減表現は (2) 式の固有値問題に帰着する．

$$\mathbf{L}\mathbf{y} = \lambda\mathbf{D}\mathbf{y} \quad (2)$$

$$\mathbf{y} = \frac{(\mathbf{D}^{\frac{1}{2}}\mathbf{x})^T(\mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}})(\mathbf{D}^{\frac{1}{2}}\mathbf{x})}{(\mathbf{D}^{\frac{1}{2}}\mathbf{x})^T(\mathbf{D}^{\frac{1}{2}}\mathbf{x})} \quad (3)$$

(2) 式の固有ベクトルが電子診療報酬データの次元削減表現であり $\mathbf{y} = [y_1, y_2, \dots, y_k]$ は類似度行列 $\mathbf{W}_{i,j}$ 、次数行列 $\mathbf{D}_i$ 、符号化した医療情報 $\mathbf{x}$ を用いることで (3) 式のように算出できる．また次元削減表現として採用する固有ベクトル数の決定は、表現の冗長および非冗長性を決定する重要な課題であり Spectral Clustering においてはクラスタ数に相当する．上記の課題において文献 [3] では Spectral Clustering のパラメータ自動設定のため多種のデータセットにおいてクラスタ数 (k) を調査している．その結果 $k = 7$ が最適なクラスタリング結果を得ると報告しているが、本研究では文献 [1] に基づき採用する固有ベクトル数が出力クラス数と等しくなるように設定する．

(b) Group Lasso

次に分類器への入力の指標となる MGAS を用いた医療情報のスコアを推定する手法について具体的に述べる．特に文献 [2] では、 $\mathbf{y}$ に対して L1 正則化である Lasso 回帰を適用することで医療情報スコア行列 $\mathbf{a}_k$ が疎行列となり各種タスクに有用なパラメータ行列を取得できることを示している．しかし、電子診療報酬データは各属性に対して複数の特徴が内包されており、Lasso 回帰で属性を選定することは難しい．例えば、電子診療報酬データの属性の一つに利用者が診断された主傷病がある．この傷病情報は 7000 種類に及ぶ．そこで本論文では各属性毎の特徴を評価するため、Group Lasso 推定を行い医療情報スコア $\mathbf{a}_k$ を算出する．(4) 式に次元縮約表現 $\mathbf{y}$ の Group Lasso 推定式を示す．

$$\arg \min_{\mathbf{a}_k} \left\| \mathbf{y}_k - \sum_k \mathbf{X}^T \mathbf{a}_k \right\|^2 + \beta \sum_k \|\mathbf{a}_k\|_2 \quad (4)$$

$$\text{MGAS}(j) = \arg \max_k |\mathbf{a}_{k,j}| \quad (5)$$

本論文では Skip-Gram を用いて医療情報を符号化したため、符号化後の各属性毎にグループ化を施した．また β は罰則項パラメータであり、本実験では (4) 式の最小二乗誤差が最小になるよう Grid Search によってパラメータを推定した．一般に Lasso の推定解は解析的に算出することが難しいため本論文では (4) 式の推定解 $\mathbf{a}_k$ の算出に、ブロック座標降下法 (Block Coordinate

(注2) : <https://code.google.com/archive/p/word2vec/>

Descent:BCD) [7] を用いる。

(c) 医療情報スコア推定

本論文では医療情報スコア $\mathbf{a}_k$ を推定する際に BCD [7] を用いる。BCD は座標降下法 (Coordinate Descent:CD) の拡張である。BCD は推定する変数を任意数のブロックに分割し、逐次処理毎に変数、定数として扱うブロックを変化させることで最適化問題を部分問題として処理する。まず Group Lasso 推定解を BCD により算出するため (4) 式を J 個のベクトル $\mathbf{a} = \{\mathbf{a}_1, \dots, \mathbf{a}_J\}$ に分割する。また上記における Group Lasso は文献 [7] より (6) 式の最適化問題へと置き換えることができる。

$$\arg \min_{\mathbf{a}_j} \frac{1}{2} \mathbf{a}_j^T M_j \mathbf{a}_j + p_j^T \mathbf{a}_j + \beta \|\mathbf{a}_j\| \quad (6)$$

$$\mathbf{a}_j^* = -(M_j + \frac{\beta}{\delta} I)^{-1} p_j \quad (7)$$

(6) 式において、 M_j は電子診療報酬データのブロック行列であり、 $M_j = X_j^T X_j$ と表現できる。また $p_j = X_j^T \mathbf{y}$ であり、(6) 式は J 個に分割した医療情報スコア行列の任意の j 番目のベクトルに対する最小化を示す。Group Lasso は縮小推定法の一種であるので、算出するパラメータの信頼領域問題を満たす最適解は非零となる。そこで最適なパラメータを $\mathbf{a}_j^*$ とする。 $\mathbf{a}_j^*$ を導出し、BCD の大域的収束性を保証するため $\|\mathbf{z}_j(\delta)\| = 1$ を満たすような δ を持つ (6) 式の最適解は (7) 式のように表すことができる。ここで $\mathbf{z}_j(\delta)$ は $\mathbf{a}_j^* = \delta \mathbf{z}_j(\delta)$ より算出し、(7) 式から (8) 式を得る。(8) 式の導入には M_j 行列を固有値分解し、固有値 γ_i および固有ベクトル q_i に分け $M_j q_i = q_i \gamma_i$ を利用した。

$$\|\mathbf{z}_j(\delta)\|^2 = \sum_i \frac{(q_i^T p_j)^2}{(\gamma_i \delta + \beta)^2} \quad (8)$$

以上によって医療情報スコア $\mathbf{a}_k$ を推定する。

MGAS の処理概要をアルゴリズム 1 に示す。なお BCD の収束条件としては、各 group 変数毎の処理回数および最小二乗誤差の閾値処理を適用した。

4.4 分類器

本論文では電子診療報酬データの多角的解析の端緒として表 1 に示した利用者の予後を表す転帰の分類を行う。転帰に関する解析処理は本論文で扱う電子診療報酬データや、より多様な情報を持つ電子カルテを対象に行われている。本論文で扱う転帰区分は表 1 の通りであるが、より広義には静的な転帰および動的な転帰に大別される。静的な転帰は各利用者の一度のみの入退院情報に関する予後を示し、動的な転帰は複数回の入退院情報を示す。例えば、動的な転帰は全データセットにおいて任意の利用者の入退院情報が複数あり、ある医院への転院後の予後も解析対象となる。本論文では静的な転帰情報を扱い、電子診療報酬データの分類器を構成するため長期間に渡って過去情報を反映可能な長短期記憶 (Long Short Term Memory:LSTM) を活用する。LSTM は RNN の発展型であり、長期記憶を可能

Algorithm 1 Multi Group Attribute Selection

```

1: Input:  $\mathbf{X} = \{x_{i,j}\}$ , parameter:  $\{\beta, \Delta\}$ 
2: Output: Medical information score matrix,  $\mathbf{a} = a_1, \dots, a_J$ 
3: Compute (a) affinity matrix,  $\mathbf{W}$  and degree matrix,  $\mathbf{D}$ 
4: Compute (b)  $\mathbf{y}$  using (a)
5: Set :  $k=1$ , initial value  $a^0 = 0$ 
6: while  $k = k_{max}$  or  $a^k$  satisfies the stopping criterion. do
7:    $a \leftarrow a^{k-1}$ 
8:   for  $j = 1, \dots, J$  do
9:      $d \leftarrow \sum_{i \neq j} \mathbf{X}_i \mathbf{a}_i - \mathbf{y}$ 
10:     $p \leftarrow \mathbf{X}_j^T d$ 
11:    if  $\|p\| \leq \beta$  then
12:       $\hat{a}_j \leftarrow 0$ 
13:    else
14:      Compute  $\|z_j(\delta)\|^2$  using Newton's method by (8)
15:      Compute  $a_j^*$  by (7)
16:    end if
17:     $a^k \leftarrow \hat{a}$ ,  $k \leftarrow k + 1$ 
18:  end for
19: end while
20: return:  $\mathbf{a} = \{a_1, \dots, a_J\}$ 

```

にしたネットワークである。LSTM の構造は、RNN の中間層を情報の破棄および保持を担う入力ゲート (Input Gate)、忘却ゲート (Forget Gate)、出力ゲート (Output Gate)、メモリセル (Memory cell) に置換したネットワークである。次に各ゲートの定義を (9), (10), (11) 式に示す。

$$g_j^{I,t} = f\left(\sum_i w_{ji}^{I,in} x_i^t + \sum_{j'} w_{jj'}^{I} z_{j'}^{t-1} + w_j^I s_j^{t-1}\right) \quad (9)$$

$$g_j^{F,t} = f\left(\sum_i w_{ji}^{F,in} x_i^t + \sum_{j'} w_{jj'}^{F} z_{j'}^{t-1} + w_j^F s_j^{t-1}\right) \quad (10)$$

$$g_j^{O,t} = f\left(\sum_i w_{ji}^{O,in} x_i^t + \sum_{j'} w_{jj'}^{O} z_{j'}^{t-1} + w_j^O s_j^t\right) \quad (11)$$

上記の (9), (10), (11) 式に示したゲートにて f はロジスティックシグモイド関数を示す。これによって各ゲートの値域は $[0, 1]$ となる。

LSTM は 3 種のゲートの値を制御する重み値を学習することによって RNN における勾配消失問題を克服している。次に LSTM の出力値である z_j^t の順伝搬計算結果を (12) 式に示す。

$$z_j^t = g_j^{O,t} f(g_j^{F,t} s_j^{t-1} + g_j^{I,t} f(u_j^t)) \quad (12)$$

(12) 式は、(1) 式に示した RNN の出力値を各ゲートによって制御していることを示している。また (12) 式から、忘却ゲート ($g_j^{F,t}$) が 0 であり入力ゲート ($g_j^{I,t}$) が 1 であれば入力値を保持し続ける RNN と同値の結果が得られることが分かる。LSTM の学習には通時間逆伝搬 (Back propagation through time:BPTT) を利用する。BPTT は RNN を順伝搬形式に変換

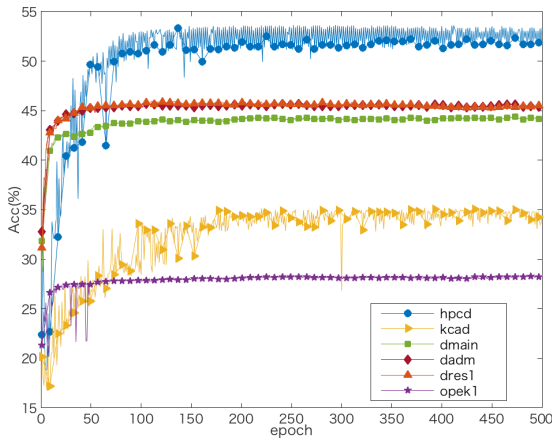


図 2 電子診療報酬データの属性毎の分類精度

した際の誤差逆伝搬法とみなすことができ、本論文では局所極小解への収束を抑制する確率的勾配降下法に上記の学習アルゴリズムを適用した。

本論文で用いたモデルである図 1 について述べる。本論文では RNN の発展系である LSTM を用いて符号化した入力情報から転帰の分類を行うためモデルの出力層にはソフトマックス関数 (Softmax Function) を分類対象クラス数分設置した。ソフトマックス関数は出力層全てのニューロンの出力値の総和が 1 の分類確率として扱えるため、最も出力値の大きいニューロンを分類結果のクラスとして判定する。また深層学習を用いたモデルでは、用途に沿った損失関数 (Error function) を設定する必要がある。そこで多クラス分類で一般的に用いられる交差エントロピー (Cross Entropy) を用いた。モデルの学習に伴い、入力データを複数の Batch に分割した Mini Batch を作成する。また本論文で用いる RNN は一般的に時系列データの処理に用いられるネットワークであるが、入力情報を同一のカテゴリに分類した後、符号化を行うことで過去情報を加味した出力結果を得た研究が報告されている [4]。本論文では符号化した入力情報をカテゴリ別に分類し Mini Batch を生成する。生成した Mini Batch を Medical Minibatch とし、生成過程を図 1 に示す。また図 1 中に示した Medical Minibatch の生成過程を事例を用いて説明する。まず各患者毎に基準とする属性によって分類するため、今回は主傷病名である dmain によって分類する。分類後に取得した他の属性情報である dmain, opek1, stay を入力、転帰を示す outcome を出力とした Batch データを学習データ及び分類精度を評価するためのテストデータに分割し出力する。また Medical Minibatch の生成に伴って 1 度のみ現れた入院情報を除外した。モデルの学習方法としては学習の進行に伴い学習係数を変化させる手法である Adagrad を適用した。また深層学習では多層に渡ってニューロンを設置するためモデルの表現能力は向上するが、一方で学習過程で計算した誤差関数の最小値が極小値である場合、学習誤差と汎化誤差が乖離した過学習となる。そのため本論文ではモデルの過学習を抑制するためニューロンを確率的に選別し自由度を抑制する手法である Dropout およびモデルの結合重みの二乗和に正則化係数を乗じた値を重みとして用いる L2 正則化を適用した。

5. 評価実験

本論文では提案モデル及び属性選択の有効性を検証するため、実データを用いた実験を行った。使用したデータセットは 33208 件の入院情報を持つ電子診療報酬データであり、テストデータに対する分類精度を正確に検証するため、4 分割交差検証 (4-fold cross validation) を用いた。また LSTM を利用した分類器にて学習率を 0.05, L2 正則化および Dropout のパラメータをそれぞれ 0.001, 0.5 とした。実験は 192GB のメモリ、Intel Xeon(R) E5-2637 3.50GHz の CPU, Quadro M5000 の GPU を搭載した Ubuntu 上のマシンで実施した。本実験は以下の諸問題に取り組む。

Q1 電子診療報酬データから転帰分類に有用な属性の検討

Q2 選択属性の分類精度検証

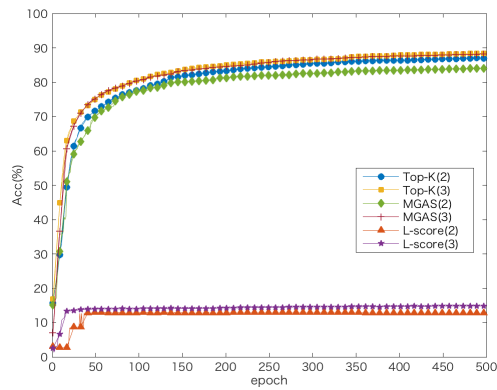
Q3 提案手法における分類精度と既存手法との比較

Q1: 転帰分類に有用な属性の検討

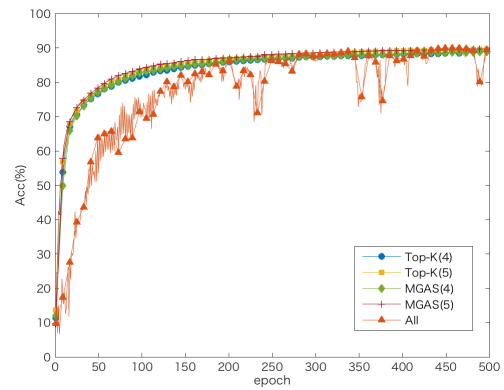
本節では、転帰分類に有用な特徴発見のため Top-K, MGAS を実データに適用し選択した複数属性を用いて分類精度を検証したケーススタディを示す。図 2 は提案モデルに転帰属性を除いた電子診療報酬データの全 23 属性をそれぞれ入力とし、MGAS にて算出した医療情報スコアの上位 6 属性の分類精度を示す。なお横軸は学習回数、縦軸は分類精度を示す。各属性について説明する。hpcd は各医療機関を識別する情報、kcad は患者が受診した診療科情報、dadm は入院契機となった傷病、dres1 は医療資源を最も消費した傷病である。図 2 において最も高精度となった属性は hpcd であることから、医療機関の利用者の予後を表す転帰は受診した医療機関によって最も区別されるという結果を得た。また属性 opek1 は MGAS にて算出したスコアでは最上位を示したが、Top-K による分類精度の高さは全属性において $\frac{9}{23}$ 番目であった。そのため属性 opek1 単体では高い分類精度は望めないが、他の属性を加えることで分類精度に影響を与える共起因子であると考えられる。

Q2: 選択属性の分類精度検証

図 3(a) は Top-K および MGAS を用いて選定した特徴を提案分類モデルへ入力した際の分類精度を示しており、括弧内の数値は利用した属性数を示す。なお L-score は MGAS にて医療情報スコアが低く、分類に有用でないと考えられる属性を利用した際の結果を示す。また図 2 に示した各属性毎の分類性能評価と異なり、図 3(a) では複数属性を分類器への入力としている。そこで 4.4 節で述べた Medical MiniBatch を属性選択法間の整合性を保持しつつ生成するため、図 2 にて最も高精度であった属性 hpcd を基準にした。各属性選択法の結果の詳細は表 2 に示す。より具体的に図 3(a) について述べる。Top-K および MGAS の精度から Wrapper 法の属性選択法では分類への有用性は低いと考えられる属性 opek1 を分類器への入力とした場合であっても同程度の分類精度を得た。そのため複数特徴間の共起影響が分類精度に関与することが確認できる。また

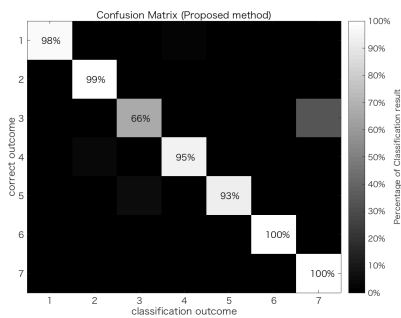


(a) 属性選択した分類精度-1(Top-K, MGAS)

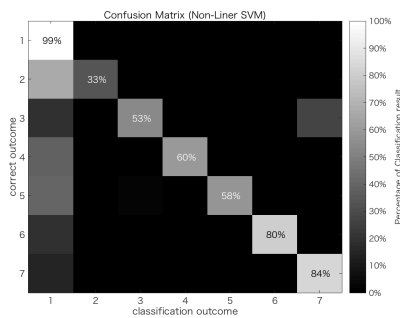


(b) 属性選択した分類精度-2(Top-K, MGAS)

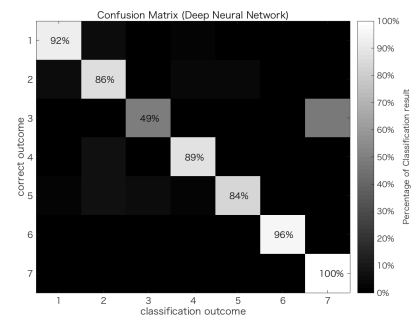
図 3 選択した属性を利用した転帰分類精度



(a) 提案モデル (入力:MGAS(5))



(b) 非線形 SVM(入力:MGAS(5))



(c) DNN(入力:MGAS(5))

図 4 提案モデルと既存モデルの比較

L-score の結果から、電子診療報酬データから任意に属性を選択するのではなく提案手法によってスコアを算出することで大幅な分類精度の向上が伺える。続いて図 3(b) について述べる。ここでは Top-K および MGAS によって利用する属性数を増加させ、電子診療報酬データの全属性を分類器への入力とした場合との比較を行った。より詳細な結果は表 2 に示した。表 2 より Filter 法の属性選択法である MGAS が Top-K を用いて選定した属性を利用した場合より精度が上回っていることが確認できる。このことから属性 opek1 の他属性との共起影響が確認できる。また全属性を入力した場合を図 3(b) 中にて All と表現した。この場合、属性選択を行った場合と同程度まで精度が向上していることが分かるが分類精度の分散が大きく、収束していない。深層学習を用いた分類器において学習器の epoch 数をテスト精度が最も高精度となる時点で停止させることはモデルの損失値や学習率から判別できない場合、実際にテストデータセットを利用しなければならない。また分類器学習時間がより多量に必要な点から属性選択の重要性ならびに MGAS の有効性を示した。

Q3: 提案手法と既存手法の比較

提案モデルの分類精度を検証するため、既存分類器との比較を行った。既存手法として採用した非線形 SVM は 3.1 節で述べたように有用な分類手法であり、カーネル関数として一般的に用いられている RBF カーネルを利用した。また提案モデルにおける LSTM の有効性を検証するため、図 1 中

表 2 属性選択法を用いた分類精度			
	Acc(%)	次元数	属性
Top-K(2)	85.43	400	hpcd, dres1
MGAS(2)	82.52	400	opek1, hpcd
L-score(2)	12.87	400	ambul, opek4
Top-K(3)	86.59	600	hpcd, dres1, dadm
MGAS(3)	86.68	600	opek1, hpcd, dmain
L-score(3)	14.55	600	ambul, opek4, dcin3
Top-K(4)	87.39	800	hpcd, dres1, dadm, dmain
MGAS(4)	87.59	800	opek1, hpcd, dmain, kcad
Top-K(5)	87.97	1000	hpcd, dres1, dadm, dmain, kcad
MGAS(5)	88.60	1000	opek1, hpcd, dmain, kcad, dadm
All Attribute	85.97	4400	All Attribute

の LSTM ニューロンを Softplus 活性化関数に置換した Deep Neural Network(以下 DNN) を利用する。Softplus 活性化関数は入力値に応じて非制限な出力値を算出するため ReLu 関数と並び、多用される活性化関数である。そして LSTM の有効性検証のため DNN は提案モデルと同様に Dropout, L2 正則化を設定した。図 4 は表 2 にて最も分類精度の高い属性選択結果である MGAS(5) を用いた場合の提案モデルと非線形 SVM, DNN における分類の精度を示している。より具体的には分類の正解クラスおよび推定したクラスに関する混同行列を作成し、マクロおよびマイクロ平均 F 値を計算した。それぞれの評価尺度は、本論文で扱う多クラス分類にて広く用いられる手法である [13]。なお評価尺度の算出において i 番目の転帰クラスにお

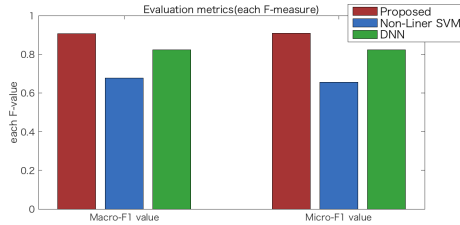


図5 Macro, Micro-F 値

ける1番目のインスタンスの分類結果を y_i^l とし ground truth を $\hat{y}_i^l$ と表現することでそれぞれ (13), (14) 式のように算出される. また $(\cdot)_+ = \max(\cdot, 0)$ である.

$$\text{Macro-F1} = \frac{1}{N} \sum_i^N \frac{2 \sum_l (y_i^l + (\hat{y}_i^l)_+)}{\sum_l (\hat{y}_i^l)_+ + \sum_l (y_i^l)_+} \quad (13)$$

$$\text{Micro-F1} = \frac{2 \sum_i^N \sum_l (y_i^l + (\hat{y}_i^l)_+)}{\sum_i^N \sum_l (y_i^l)_+ + \sum_i^N \sum_l (\hat{y}_i^l)_+} \quad (14)$$

正確にクラスを推定することができた場合, 混同行列は対角行列となり, 各 F 値は 1 となる. 図 4(a), 図 4(b), 図 4(c) はそれぞれ提案モデルと非線形 SVM, DNN の混同行列を示しており, 内部数値は各クラスの正答割合を示す. 図 4(a) から提案モデルの分類器は, 多くの転帰クラスにおいて高精度の分類が実現できていることが分かる. 特に図 4(b) と比較して転帰クラス 2 および 4~7 の大幅な精度改善が確認できる. また図 4(c) に着目すると, 各転帰クラスにおいて提案モデルがより高精度に分類できており提案モデルによって他利用者の過去情報を加味したことが高精度に起因したと考えられる. 特筆すべき点として, 各手法において転帰クラス 3 の分類精度が低い. 転帰クラス 3 は傷病が寛解したと判断される場合を示しており, 図 4 より転帰クラス 7 への誤分類が多いことが分かる. 上記の問題に対して, 今回実験に使用したデータセットを精査したところ転帰クラス 3 に該当する患者の 71.1% が ICD10 の階層分類において, がん等の悪性新生物の疾患が含まれる階層であることが判明した. 一般にがん患者は複数の併存症を患っている割合が高く, 臓器機能が低下していることも多いため, がん治療により合併症発症のリスクの増加および副作用が遷延しやすい傾向がある. そのため, 傷病以外による死亡を示す転帰クラス 7 と類する入院エピソードを持った患者が転帰クラス 3 においても多く存在し分類が困難になったと考えられる. また統合的な評価尺度として図 5 にマクロおよびマイクロ平均 F 値を示した. 図 5 より提案モデルは非線形 SVM, DNN よりも正確な転帰分類を実現していることが分かる.

6. む す び

本論文では電子診療報酬データの多角的解析のための医療情報分類手法として医療情報から分類に有用な特徴を特定し, 深層学習を用いた分類モデルを構築した. 提案手法は, 既存の属性選択法では特定の難しい分類に有用な特徴の選定が可能であり, 実データを用いて転帰分類に関する実験を行った結果, 既存モデルと比較して提案モデルの分類精度の向上を示した.

文 献

- [1] Deng Cai, Chiyuan Zhang, Xiaofei He. Unsupervised Feature Selection for Multi-Cluster Data, *ACM SIGKDD Conference on Knowledge Discovery and Data Mining(KDD)*, pp.333-342(2010)
- [2] Xiaofei He, Deng Cai, Partha Niyogi. Laplacian score for feature selection. *NIPS'05 Proceedings of the 18th International Conference on Neural Information Processing Systems*, pages 507-514, 2005
- [3] Lihi Zelnik-Manor, Pietro Perona. Self-tuning spectral clustering. *NIPS'04 Proceedings of the 17th International Conference on Neural Information Processing Systems*, pages 1601-1608, 2004
- [4] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, Domonkos Tikk. Session-Based Recommendations with Recurrent Neural Networks, *arXiv:1511.06939v4, ICLR2016*
- [5] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. *AAAI'16 Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3697-3703, 2016
- [6] M.Yuan and Y.Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society:Statistical Methodology*, pp49-67(2006)
- [7] Zhiwei Qin, Katya Scheinberg, Donald Goldfarb. Efficient block-coordinate descent algorithms for the Group Lasso. *Mathematical Programming Computation, Volume5*, pp143-169(2013)
- [8] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao. Learning to read chest X-rays:recurrent neural cascade model for automated image annotation. *Conference on Computer Vision and Pattern Recognition(CVPR)*, pp2497-2506(2016)
- [9] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *In Proceedings of Workshop at International Conference on Learning Representations 2013:ICLR2013*
- [10] Nozomi Nori, Hisashi Kashima, Kazuto Yamashita, Hiroshi Ikai, Yuichi Imanaka. "Simultaneous Modeling of Multiple Diseases for Mortality Prediction in Acute Hospital Care". *KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855-864, 2015
- [11] Naren Meadem, Nele Verbiest, Kiyana Zolfaghar, Jayshree Agarwa, Si-Chi Chin, Senjuti Basu Roy. Exploring Pre-processing Techniques for Prediction of Risk of Readmission for Congestive Heart Failure Patients. *KDD-DMH'13 Workshop on Data Mining for Healthcare*, 2013
- [12] Hwanjo Yu, Jadeo Vaidya, Xiaoqian Jiang. Privacy-Preserving SVM Classification on Vertically Partioned Data. *PAKDD'06 Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 647-656, 2006
- [13] Yuhong Guo, Wei Xue. Probabilistic multi-label classification with sparse feature learning. *IJCAI '13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1373-1379, 2013

謝 辞

本研究の一部は JSPS 科研費 JP15H02705, JP17H04681, JP16K12430, JST さきがけ, 総務省 SCOPE (受付番号 162110003), 厚生労働科学研究費補助金 (H29-ICT-一般-007) 及び国立研究開発法人日本医療研究開発機構 (AMED) の臨床研究等 ICT 基盤構築研究事業の助成を受けたものです.