(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6468652号
(P6468652)
(45)発行日　平成31年2月13日(2019.2.13)　(24)登録日　平成31年1月25日(2019.1.25)

(51)Int.Cl.　　F I
G06F 16/00 (2019.01)　　G06F 17/30 220A
G06F 16/30 (2019.01)　　G06F 17/30 170A

請求項の数 9　(全 23 頁)

(21)出願番号　特願2015-144327(P2015-144327)
(22)出願日　平成27年7月21日(2015.7.21)
(65)公開番号　特開2017-27307(P2017-27307A)
(43)公開日　平成29年2月2日(2017.2.2)
審査請求日　平成30年2月13日(2018.2.13)

(73)特許権者　000208891
ＫＤＤＩ株式会社
東京都新宿区西新宿二丁目３番２号
(74)代理人　100092772
弁理士　阪本　清孝
(74)代理人　100119688
弁理士　田邊　壽二
(72)発明者　小川　圭介
東京都新宿区西新宿二丁目３番２号　ＫＤＤＩ株式会社内
(72)発明者　橋本　真幸
埼玉県ふじみ野市大原二丁目１番１５号　株式会社ＫＤＤＩ研究所内

最終頁に続く

(54)【発明の名称】医療データ解析装置

(57)【特許請求の範囲】
【請求項１】
異なる種類のデータが頻度要素として混在するバグオブワードの形で与えられた一連の対象者の一連の年代における医療データを、対象者及び年代ごとの個別データの集まりとして、深層学習を適用し、各中間層における当該個別データの表現データを、前記異なる種類のデータが頻度要素として混在するバグオブワードの圧縮表現として出力するデータ表現学習部を備えることを特徴とする医療データ解析装置。
【請求項２】
前記データ表現学習部は、各中間層のノード数を、層が深くなるほど減らすように設定したうえで深層学習を適用することを特徴とする請求項１に記載の医療データ解析装置。
【請求項３】
前記出力された中間層ごとの表現データを潜在トピック分析によりクラスタリングし、当該クラスタリング結果における各クラスタとクラスタ間の遷移確率とを健康状態の予測モデルとして出力するモデル構築部をさらに備えることを特徴とする請求項１または２に記載の医療データ解析装置。
【請求項４】
前記モデル構築部では、所定のクラスタ数候補ごとにクラスタリングを行い、得られたクラスタリング結果の各クラスタにおいて属する個別データが、所定の健康状態に関する評価が良いか悪いかをカウントしたクロス集計表を作成し、当該クロス集計表により求まる情報量基準の値に基づいて、前記所定のクラスタ数候補の中から最適クラスタ数を決定

し、当該最適クラスタ数におけるクラスタリング結果に基づいて前記予測モデルを出力することを特徴とする請求項３に記載の医療データ解析装置。
【請求項５】
前記データ表現学習部は、各中間層のノード数を、層が深くなるほど減らすように設定したうえで深層学習を適用し、
前記モデル構築部は、層がより深い側の中間層の表現データより出力される予測モデルを、より長期の予測を行うための予測モデルとして出力することを特徴とする請求項３または４に記載の医療データ解析装置。
【請求項６】
健康状態の予測モデルとしていずれの中間層に対応するものを利用するかと、バグオブワードの形での予測対象者の健康状態データと、当該予測対象者の現在年代と、当該予測対象者の予測対象未来年代と、の指定をユーザより受け付け、
当該指定された健康状態の予測モデル上で、当該指定された予測対象者の健康状態の、当該指定された現在年代から当該指定された予測対象未来年代に至るまでの推移を予測する予測部をさらに備えることを特徴とする請求項３ないし５のいずれかに記載の医療データ解析装置。
【請求項７】
前記データ表現学習部は、各中間層のノード数を、層が深くなるほど減らすように設定したうえで深層学習を適用し、
前記予測部は、前記推移を予測した結果が、予測精度がないと判定される場合には、より深い中間層における健康状態の予測モデルを用いて予測を行うことを特徴とする請求項６に記載の医療データ解析装置。
【請求項８】
健康状態の予測モデルとしていずれの中間層に対応するものを利用するかと、バグオブワードの形での予測対象者の健康状態データと、の指定をユーザより受け付け、
当該指定された予測モデルにおけるクラスタリング結果の各クラスタを、所定の健康状態に関する評価が良いか悪いかで２つのクラスタ群に分け、前記指定された予測対象者の健康状態データが当該２つのクラスタ群のいずれに所属するかを判定する予測部をさらに備えることを特徴とする請求項３ないし５のいずれかに記載の医療データ解析装置。
【請求項９】
前記医療データにおいて混在する異なる種類のデータは、単位系及び／又は構造が異なることで異なる種類のデータとして構成されていることを特徴とする請求項１ないし８のいずれかに記載の医療データ解析装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、異なる種類のデータ（健診値と問診結果など）が混在する医療データであっても、潜在トピック分析により適切なクラスタリングを行うことが可能となるようなデータ表現を得ることのできる医療データ解析装置に関する。
【背景技術】
【０００２】
医療データに基づいて、対象者をクラスタリングしたい場合がある。特許文献１や特許文献２に代表されるように、健康管理システム等が大きな広がりを見せている。このような健康管理システムでは、利用者に対して健康上のアドバイス等を行う場合が多いが、特許文献３に示すように、利用者を実際の健康データを元に分類した上でアドバイスを行った方が、より行動変容につながりやすい。
【先行技術文献】
【特許文献】
【０００３】
【特許文献１】特開2013-085626号公報

【特許文献２】特開2010-264088号公報
【特許文献３】特開2010-170534号公報
【発明の概要】
【発明が解決しようとする課題】
【０００４】
近年、潜在的ディリクレ配分法（Latent dirichlet allocation:LDA）に代表される高精度な分類手法として、潜在トピック分析が注目を浴びている。また、このLDAを時系列的な影響を加味するように拡張したトピックトラッキングモデル(Topic Tracking Model（TTM）)なども提案されている。このTTMを用いれば、長期の時系列的な影響を加味してクラスタリングが可能であり、長期にわたる変化のモデル等を作成することができる。
【０００５】
ここで、LDAを用いた健康予測モデルを構成するにあたり、以下の問題がある。
【０００６】
一般的に、データによる分類・予測を行う際には、様々な種類のデータが存在した方が精度の向上が見込まれる。しかしながら、LDAではデータを単語の頻度表現で表さなければならないため、単位系の異なるデータを用意した際の表現が難しいという問題がある。例えば、レセプトデータから得られた頻度表現が(糖尿病、高血圧、高脂血症)=(3,4,10)であり、レセプトデータとは別途であって単位系が異なっている健診データにおける血糖値が120であったときに、血糖値をレセプトデータと並列な形の頻度表現(糖尿病、高血圧、高脂血症、血糖値)=(3,4,10,F)に直すのは非常に難しい。
【０００７】
具体的に、上記の「血糖値=120」の「頻度表現=F」を決定するには、次のようなことを考慮しなければならない。例えば、最大値を12する値に離散化すればよいのか、100を最大値とする離散値にすればよいか、もしくは血糖値と糖尿病は関連するから、２つを統合した表現が必要か・・・といった点を考慮しなければいけない。
【０００８】
上記のように、従来技術には次のような課題があった。すなわち、頻度表現（バグオブワーズ(bag of words)表現）のように単一のデータ表現を用いてクラスタリングを行うLDA等を使って、異なる単位系が混在するデータ（例：文書やレセプトデータから得られた、特定の単語の頻度表現と、健診データのような連続値データ、問診結果のようなアンケート結果等）を分類しようとする際に、どのような単一のデータ表現を用いればよいのかという点が必ずしも明らかではなかった。
【０００９】
本発明は、上記の従来技術の課題に鑑み、適切な単一のデータ表現を得ることのできる医療データ解析装置を提供することを目的とする。
【課題を解決するための手段】
【００１０】
上記目的を達成するため、本発明は、異なる種類のデータが頻度要素として混在するバグオブワードの形で与えられた一連の対象者の一連の年代における医療データを、対象者及び年代ごとの個別データの集まりとして、深層学習を適用し、各中間層における当該個別データの表現データを、前記異なる種類のデータが頻度要素として混在するバグオブワードの圧縮表現として出力するデータ表現学習部を備えることを特徴とする。
【発明の効果】
【００１１】
本発明によれば、異なる種類のデータが混在するバグオブワード形式の医療データを、深層学習における各中間層の形に変換することで、異なる種類のデータ同士の関係が自動で適切に抽象化され圧縮された表現データを得ることができる。
【図面の簡単な説明】
【００１２】
【図１】一実施形態に係る医療データ分析装置の機能ブロック図である。

【図２】一実施形態に係る予測部の機能ブロック図である。
【図３】一実施形態に係る予測部の機能ブロック図である。
【図４】文書化部に入力される全医療データの模式的な例を示す図である。
【図５】データ表現学習部による処理を模式的に示す図である。
【図６】深層学習における各中間層の学習過程を模式的に示す図である。
【図７】クラスタリング部におけるクラスタリング結果が行列分解の形で得られることを説明するための図である。
【図８】クラスタリング部におけるクラスタリング結果及び対応する遷移確率の算出の例を[1]～[5]と分けて示す図である。
【図９】クラスタリング部におけるクラスタリング結果を健康状態の推移モデルとして利用する例を示す図である。
【図１０】AIC計算部の作成するクロス集計表を示す図である。
【図１１】分類評価部の処理内容を説明するための図である。
【図１２】分類評価部においてF値を算出する際に作成する集計表を示す図である。
【図１３】本発明の効果の実例を示す図である。
【図１４】図１０，図１１等のクロス集計表を一般化した表である。
【図１５】図１４のクロス集計表に対応する従属モデルにおける確率の表である。
【図１６】図１５のクロス集計表に対応する独立モデルにおける確率の表である。
【発明を実施するための形態】
【００１３】
図１は、一実施形態に係る医療データ解析装置の機能ブロック図である。医療データ解析装置10は、文書化部1、第一正規化部2、データ表現学習部3、第二正規化部4、モデル構築部5及び予測部6を備える。ここで、データ表現学習部3は、入力設定部31、ネットワーク重み最適化部32及び中間層出力取得部33を備える。モデル構築部5は、クラスタリング部51、AIC計算部52及び最適クラスタ数決定部53を備える。
【００１４】
図２及び図３はそれぞれ、一実施形態に係る予測部6の機能ブロック図である。図２の実施形態では、予測部6は分類評価部61及び危険判定部62を備える。図３の実施形態では、予測部6は遷移予測部65及び期間設定部66を備える。
【００１５】
以下、図１の各部の処理を説明する。
【００１６】
文書化部1では、医療データ解析装置10によるモデル構築のための入力データとしての全医療データを読み込み、当該全データを構成する各対象者Xの各年代n（年齢n）における文書化された医療データD(X, n)を生成して第一正規化部2へと出力する。
【００１７】
当該医療データD(X, n)への文書化とは、周知のバグオブワード(bag of words)の形式、すなわち所定の各単語の頻度（出現回数）を要素とする文書ベクトルの形式へ変換することであり、データD(X, n)は対象者Xのn歳時点での健康状態を反映したベクトルとなっている。当該文書化は具体的には以下の通りであり、前述の従来技術の課題が解決されていない形で文書化が行われる。すなわち、文書化部1による文書化においては、元の医療データに異なる単位系のデータが混在していることに対する特別な対処は行われずに、異なる単位系ごとにそれぞれルールベースで頻度を求めることで１つのバグオブワードを得る。
【００１８】
まず、入力される全医療データは、一連の対象者の一連の時期における健康状態を評価したものであり、具体的には例えば健康組合等のもとで実施される健康診断結果や、医師による問診の結果、あるいはレセプト（診療報酬明細書）等やこれらの組み合わせを用いることができる。ここで、本発明においては特に、異なる単位系のデータが混在する形で入力される医療データが構成されている。

【００１９】
あらかじめ、当該医療データに記載されている、あるいは、記載されうることが既知の健康状態を表す所定の複数m個の単語i_1，i_2，…，i_mを用意しておき、文書化部1において対象者Xのn歳における医療データのテキストを解析することで、単語i_1，i_2，…，i_mの頻度ベクトルとして健康状態を表すベクトルD(X, n)を生成することができる。
【００２０】
例えば、問診データ等における特定の疾病の名称に相当する単語i_bが対象者Xのn歳の医療データに存在すれば、ベクトルD(X, n)の当該i_bの要素の値を「1」とし、存在しなければ同要素の値を「0」とすることができる。レセプトデータ等における処方された薬剤名などの単語i_bについても同様に当該単語が存在するか否かで「1」または「0」とすることができる。また、同単語ibが問診データ等に複数回現れている場合は現れた回数分の要素の値としてもよいし、以下に説明する数値評価項目等の場合と同様に当該現れた回数に所定関数を適用した値を要素の値としてもよい。
【００２１】
また、健康診断データにおける体重や血液検査の結果等、数値で評価される項目については当該項目に応じた所定の単語を用意しておき、評価数値に応じた所定規則（所定関数等）により当該単語の頻度を算出してベクトルD(X,n)の要素の値とすることができる。このような評価数値から単語頻度への変換に関しては、本出願人による特開２０１５－３２０１３号公報（発明の名称：数値データ解析装置及びプログラム）、特願２０１３－１６３２０７号（数値データ解析装置及びプログラム）、特願２０１３－２１７８１７号（数値データ解析装置及びプログラム）を利用してもよい。
【００２２】
なお、上記のような数値（量的データ）の場合の他、質的データ（例えば、問診票等に記載された喫煙習慣の有無など）の場合も、同様に所定規則により対応する単語の頻度へと変換し、ベクトルD(X,n)の要素の値とすることができる。
【００２３】
以上のように、単語i_1，i_2，…，i_mの各々は、入力される医療データにおける健康状態の評価項目（直接的に評価するもののみではなく、レセプトデータにおける薬剤名のように健康状態を間接的に反映する項目も含む）の各々に対応する単語であり、対象者Xのn歳における当該評価結果に対して所定規則（単語i_1，i_2，…，i_mの各々に個別規則を用意しておくことができる）を適用することで、文書化部1では文書ベクトルD(X, n)を生成する。
【００２４】
図４に、文書化部1に入力される全医療データの模式的な例を示す。当該例に示すように、入力としての全医療データには欠損があることが多い。すなわち、健康予想モデル構築を精度よく実施するには、各対象者につき数十年等の長期間に渡るデータが存在していることが望まれるが、実際には図２の例のように、数年の短期間に渡るデータしか利用できないということが多い。
【００２５】
なお、図４の例では、例えばAさんに関しては40歳～43歳のデータが存在しているので、文書化部1においてAさんの医療データよりD(A,40),D(A,41),D(A,42),D(D43)という4個のデータが出力されることとなる。Gさん、Dさんといったその他の対象者についても同様に医療データが存在する年代分のデータが出力されることとなる。
【００２６】
以上のように、文書化部1の出力するバグオブワードD(X,n)は、異なる単位系にあるデータ同士の関係を特に考慮することなく、各要素（各単語の頻度）につき個別のルールベースで求められたものである。このため、当該バグオブワードD(X,n)に対してそのままの形でLDA等を適用しても、高精度な分類結果が得られるとは限らない。
【００２７】
このことに対する解決策を提供するのが、次の各部2,3,4（特にデータ表現学習部3）で

ある。当該各部2,3,4の処理を経ることで単位系が異なっていることを考慮したバグオブワードが得られ、モデル構築部5のクラスタリング部51でLDA等のクラスタリングの対象となる。

【００２８】

なお、本発明においてデータにおける「異なる単位系」とは、上記の文書化部1の説明より明らかなように、量的データと質的データとの区別に加え、量的データにおける種類の区別（身長と体重と血圧との区別など）及び性質の区別（離散、連続の区別や間隔尺度、比例尺度の区別など）と、質的データにおける種類の区別（特定項目の問診回答内容とレセプトに記載の特定薬剤名称との区別など）及び性質の区別（名義尺度、順序尺度の区別など）と、をも含んで、データ同士の単位系が異なっていることを意味する。

【００２９】

上記において性質の区別には、構造の区別も含まれる。例えば、HbA1cが6.5以上ならば糖尿病と判定される場合には、「医師による判定フロー」という構造が存在している。

【００３０】

第一正規化部2では、文書化部1の出力した各バグオブワードD(X,n)をそのノルム|D(X,n)|で割って正規化（各要素の値が0以上1以下になるよう正規化）し、正規化されたD(X,n)/|D(X,n)|をデータ表現学習部3の入力設定部31へと出力する。なお、以下のデータ表現学習部3の説明においては、表現の簡略化のため「正規化されたD(X,n)」あるいは単に「データD(X,n)」等の表現で、上記のように第一正規化部2が出力した「D(X,n)/|D(X,n)|」を意味するものとする。

【００３１】

データ表現学習部3は、第一正規化部2より得られる一連の正規化されたD(X,n)をもとに多層の深層学習を行うことで、当該各データD(X,n)の深層学習の層構造における各中間層の表現を得て、第二正規化部4へと出力する。ここで、深層学習(Deep Learning)は周知であり、以下の非特許文献１等に開示されている。

[非特許文献１] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18.7 (2006): 1527-1554.

【００３２】

図５は、データ表現学習部3による処理を模式的に示す図であり、深層学習による学習によって構築される層構造を示している。データ表現学習部3では、ネットワーク重み最適化部32においてこのような層構造を学習し、中間層出力取得部33においてデータD(X,n)の各中間層における表現を得て、第二正規化部4へと出力する。

【００３３】

すなわち、ネットワーク重み最適化部32では各層間におけるノード間の以下の式(1)のような関数関係を定めるためのネットワーク重みを学習する。ここで、x_iは浅い側（入力側）の層の各ノードの値であり、y_jはx_iよりも１層だけ深い側の層の各ノードの値である。ネットワーク重み最適化部32は、各層間での以下の式(1A)における重みw_{ij}及び係数b_j（すなわち、ネットワーク重み）を学習により求める。関数fにはシグモイド関数などを用いることができる。

【００３４】

【数１】

$$y_j = b_j + \sum_i f(x_i) w_{ij} \qquad \cdots (1A)$$

【００３５】

なお、上記の式(1A)のような関係は、深層学習等の分野において周知のように、図５ではノード間のエッジとして表現されている。

【００３６】

また、ネットワーク重み最適化部32で上記のようなネットワーク重みの学習を行うに際

しては、浅い側から順次、まず、図５に示す[0],[1]の間すなわち入力層と第一中間層とにおける式(1A)のネットワーク重みを学習し、次に[1],[2]の間すなわち第一中間層と第二中間層における式(1A)のネットワーク重みを学習し、次に[2],[3]の間すなわち第二中間層と第三中間層における式(1A)のネットワーク重みを学習し、…といったように学習を行うこととなる。当該学習の各段階において必要なデータをネットワーク重み最適化部32に提供するのが入力設定部31である。ここで、図１中に線L1として示すように、初回の学習においては第一正規化部2の出力した各データD(X,n)がネットワーク重み最適化部32に渡され、２回目以降の学習においては線L2として示すように、ネットワーク重み最適化部32が学習して得たネットワーク重みを用いて中間層出力取得部33が得た各データを再帰的にネットワーク重み最適化部32へと渡す。

【００３７】

中間層出力取得部33は、ネットワーク重み最適化部32の計算したネットワーク重みに基づき、データD(X,n)の各中間層における表現を求めて、データ表現学習部3における出力として第二正規化部4へと出力すると共に、上記のようなより深い側の層間のネットワーク重みの再帰的な学習を継続すべく、入力設定部31へも出力する。ここで、K番目の中間層(K=1,2,3, …,M)においてはデータD(X,n)がデータ$D(X,n)_{[K]}$として表現されているものとする。

【００３８】

すなわち、図５では[0]の入力層にデータD(X,n)を入力することで、中間層出力取得部33は[1]の第一中間層におけるデータ$D(X,n)_{[1]}$を得る。当該データ$D(X,n)_{[1]}$は[1-1]に示すように、後段側の第二正規化部4及びモデル構築部5に渡され、第一モデルの生成に利用される。以降も同様の処理が深い層に向かって継続される。すなわち、図５の[1]の第一中間層にそのデータ$D(X,n)_{[1]}$を入力することで、中間層出力取得部33は[2]の第二中間層におけるデータ$D(X,n)_{[2]}$を得る。当該データ$D(X,n)_{[2]}$は[2-1]に示すように、後段側の第二正規化部4及びモデル構築部5に渡され、第二モデルの生成に利用される。[3],[3-1]に示すように同様に、第三中間層における$D(X,n)_{[3]}$を得て、第三モデルの生成に利用する。

【００３９】

本発明においては、データ$D(X,n)_{[K]}$は、当初のデータD(X,n)が深層学習により抽象化されたものとして得られることを利用して、当初のデータD(X,n)における単位等の異なるデータ由来の要素の混在の問題を解決する。すなわち、深層学習の深い層に進むにつれ、当初のデータD(X,n)がより抽象化され、血液検査値とレセプト評価記載といったような、性質の異なるデータ同士の関係が自動的に適切に表現された形で、データ$D(X,n)_{[K]}$が得られる。つまり、異なる単位系や異なる構造の存在するデータD(X,n)を単一の表現に圧縮したものとして、データ$D(X,n)_{[K]}$が得られる。当該データ$D(X,n)_{[K]}$よりモデル構築部5において高精度な予測モデルを構築することが可能となる。

【００４０】

なお、ネットワーク重み最適化部32でネットワーク重みを求める対象となる、図５のようなネットワーク構造はユーザ等が予め与えておく。まず、[0]の入力層のN0個の各ノード0-1～0-N0は、当初のデータD(X,n)の各要素に対応するものとして与えておく。（すなわち、データD(X,n)の次元数が入力層のノード数N0と一致する。）さらに、第一中間層のノード1-1～1-N1におけるノード数N1、第二中間層のノード2-1～2-N2におけるノード数N2、第三中間層のノード3-1～3-N3におけるノード数N3、…といったような、K(K=1,2,3, …, M)番目の中間層のノード数NKについても、ユーザがハイパーパラメータとして事前に与えておくことができる。

【００４１】

ここで、ノード数NKは、より深い中間層に行くほどそのノード数が減るように、以下のような関係で与えるようにすることが好ましい。

N0＞N1＞N2＞N3＞…＞NK＞NK+1＞…＞NM

【００４２】

図６は、周知事項である深層学習における各中間層の学習を模式的に説明する図であり、図５のネットワーク構造を学習する場合の例が示されている。図６にて[1]は第一中間層の学習を示しており、[1-1]に示す入力層と、[1-2]に示す第一中間層と、[1-3]に示す仮の出力層と、のネットワーク構造を設けたうえで、入力層のデータができる限り同じ形で、仮の出力層のデータとして得られるように、第一中間層の学習が行われる。この際、オートエンコーダ(Auto-encoder)あるいは制限付きボルツマンマシン(RBM; Restricted Boltzmann Machine)等を用いて、誤差又はエネルギーを最小化するようにすればよい。

【００４３】

図６にて線L12に示すように、学習された第一中間層のデータ表現を新たな入力及び出力として設定することで、[2]に示すように第二中間層の学習が行われる。[2-1]が第一中間層の入力、[2-2]が第二中間層、[2-3]が仮の第一中間層の出力であり、[1]と同様に[2-1]の入力が[2-3]の出力において可能な限り再現されるように、[2-2]の第二中間層が学習される。

【００４４】

同様に、図６にて線L23に示すように、学習された第二中間層のデータ表現を新たな入力及び出力として設定することで、[3]に示すように第三中間層の学習が行われる。[3-1]が第二中間層の入力、[3-2]が第三中間層、[3-3]が仮の第二中間層の出力であり、[1],[2]と同様に[3-1]の入力が[3-3]の出力において可能な限り再現されるように、[3-2]の第三中間層が学習される。図６には示していないが、より深い中間層の学習も全く同様である。

【００４５】

なお、図６の線L12,L23といったデータ設定を担うのが、入力設定部31である。

【００４６】

深層学習においては、図６のような各中間層の学習を行い、最終的な出力層まで到達した時点で、再度、全体的なチューニングに相当する学習が行われる。本発明においては、データ表現学習部3では当該全体的なチューニングに相当する学習を行ってもよいし、これを省略して中間層の部分の学習のみを用いるようにしてもよい。

【００４７】

第二正規化部4は、データ表現学習部3より得られた各中間層KのデータD(X,n)$_{[K]}$を正規化して、モデル構築部5のクラスタリング部51へと出力する。当該正規化とは、データD(X,n)$_{[K]}$の各要素が例えば0～1の範囲にあるのを、LDA等のクラスタリングを行う対象としてのバグオブワード形式（単語頻度形式）にすることであり、例えば所定の最大単語数を乗じて小数点以下は四捨五入等することで、正規化することができる。

【００４８】

ここで、第一正規化部2が出力するデータD(X,n)に関してデータ表現学習部3で参照される際の用語の使い方を定義したのと同様に、第二正規化部4が出力する正規化されたデータD(X,n)$_{[K]}$についても、モデル構築部5で参照する場合には、「正規化されたデータD(X,n)$_{[K]}$」あるいは単に「データD(X,n)$_{[K]}$」で、正規化されていることを表すものとする。

【００４９】

モデル構築部5では、図５の[1-1],[2-1],[3-1]で説明したように、第二正規化部4が出力するK番目の中間層ごとの正規化された一連のデータD(X,n)$_{[K]}$をクラスタリングして、それぞれ、クラスタリング結果CL(K)を得ると共に、当該クラスタリング結果CL(K)より健康状態等に関する予測モデルPM(K)を生成する。図１に示すように、生成された予測モデルPM(K)はユーザの参照に供すべく出力されると共に、予測部6へと渡される。

【００５０】

このような処理を行うに際して、モデル構築部5の各部51～53は以下のように機能する。

【００５１】

クラスタリング部51は、一連のデータD(X,n)$_{[K]}$を指定された一連のクラスタ数mのそれぞれの値のもとでクラスタリングして、各クラスタ数mにおけるクラスタリング結果CL(K)

$_{[m]}$を得て、当該結果をAIC計算部52へと渡す。AIC計算部52は、当該クラスタリング結果CL(K)$_{[m]}$よりクロス集計表を作成して、そのAIC値（赤池情報量基準の値）（AIC(K,m)とする）を計算し、当該値AIC(K,m)を最適クラスタ数決定部53に渡す。

【００５２】

最適クラスタ数決定部53は、K番目の中間層のデータD(X,n)$_{[K]}$より生成される予測モデルPM(K)を、値AIC(K,m)が最小になるようなクラスタ数m=m$_{[最小]}$（すなわち、最適のクラスタ数）におけるクラスタリング結果CL(K)$_{[m[最小]]}$によって生成し、ユーザ及び予測部6へと出力する。

【００５３】

以下、各部51,52の詳細を説明する。なお、クラスタリング部51にてクラスタリングを行う際のクラスタ数mが、一連の候補の値が与えられているのではなく、ユーザ指示等によって事前に１通りの値のみに限定されていれば、AIC計算部52及び最適クラスタ数決定部53を省略してもよい。この場合、クラスタリング部51において当該１種類のクラスタ数mでクラスタリングを行い、結果CL(K)を得ると共に予測モデルPM(K)を生成すればよい。

【００５４】

クラスタリング部51は、まず、クラスタ数mのもとで一連のデータD(X,n)$_{[K]}$をクラスタリングしてクラスタリング結果CL(K)$_{[m]}$を得る。当該クラスタリングにはLDA等の潜在トピック分析の手法を用いることができる。LDAについては以下の非特許文献２等に開示されている。

[非特許文献２] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research,3:993-1022, January 2003.

【００５５】

図７は、クラスタリング部51におけるLDA等の潜在トピック分析の手法によるクラスタリングの結果が行列分解の形で得られることを説明するための図である。

【００５６】

図７に示すように、LDA等の潜在トピック分析では分類対象の全データD（K番目の中間層ごとの一連のデータ｛D(X,n)$_{[K]}$｝。図７ではKに依存する旨の表示は省略している。）は単語iの頻度ベクトルとして与えられている各文書u（本発明では文書化部1の各データD(X,n)の(X,n)すなわち特定対象者Xの特定年代nのデータに相当する。）からなる。なお、中間層からの出力データD(X,n)$_{[K]}$は前述のようにデータの縮約が施された状態にあるので、当初のデータD(X,n)にあったような明示的な単語iが存在するわけではないが、本発明においては特別に、このようなデータD(X,n)$_{[K]}$を単語頻度として扱っている。

【００５７】

そして、図７に示すように、当該全データDにクラスタリングを行った結果が、文書uとトピックkとの関係を表すθ行列とトピックkと単語iとの関係を表すΦ行列との行列としての積「D=θ×Φ」として得られ、クラスタリング部51では当該行列分解結果を出力する。

【００５８】

当該行列分解結果において、各トピックkが各クラスタに対応するものとすると、文書uのトピック比率を表すθ行列の各行は、各文書uのクラスタ所属確率と解釈できる。当該クラスタ所属確率は、各文書uにおける各トピックkのトピック比率であり、対応する元のデータD(X,n)$_{[K]}$の健康状態を表現したベクトルとなっている。従って例えば、各文書u(=各データD(X,n)における(X,n))は、その最大のトピック比率の値のトピックに対応するクラスタに所属しているものとして、クラスタリング結果を解釈することができる。

【００５９】

クラスタリング部51は、さらに、クラスタリング結果CL(K)$_{[m]}$を予測部6において予測モデル（の一実施形態）として利用可能なように、当該結果CL(K)$_{[m]}$における各クラスタ間の遷移確率を計算する。（なお、遷移確率の計算は、最適クラスタ数決定部53において最適なクラスタリング結果を与えるものとして決定された最適クラスタ数m$_{[最小]}$についてのみ実施するようにしてもよい。）

【００６０】
ここで説明のため、クラスタリング結果の各クラスタをCi(i=1, 2, …)と書くことにする（K依存の旨は表記が煩雑となるため省略する。）と、クラスタCi,Cj間の遷移確率P(Ci→Cj)は、以下の第一方針及び第二方針で定まる一連のカウントを集計し、クラスタ毎に確率として規格化することにより、クラスタリング部51において算出すればよい。
【００６１】
すなわち、第一方針として、隣接する年代n歳及びn+1歳についての同じ対象者Xのデータ D(X,n),D(X,n+1)が異なるクラスタCi,Cj(i≠j)に分類されている場合、すなわち、D(X,n)∈CiかつD(X,n+1)∈Cjである場合、クラスタCiよりクラスタCjへと至る遷移が1回あったものとしてカウントすることにする。（ここで、遷移の方向Ci→Cjは、対象者Xのn歳時点での所属クラスタCiからn+1歳時点での所属クラスタCjへと向かう方向、すなわち年代の進む方向である。）
【００６２】
また、第二方針として、隣接する年代n歳及びn+1歳についての同じ対象者Xのデータ D(X,n),D(X,n+1)が同じクラスタCiに分類されている場合、すなわち、D(X,n)∈CiかつD(X,n+1)∈Ciである場合、クラスタCiよりクラスタCi自身へと至る遷移（クラスタCiの自己遷移）が1回あったものとしてカウントすることにする。（なお、第一方針でi=jとした場合が第二方針である。）
【００６３】
以上、第一、第二方針より、以下の式(1)のようにクラスタCi,Cj間の遷移確率P(Ci→Cj)をクラスタCiからCjへの遷移数N(Ci→Cj)に比例するように計算することができる。第一方針(i≠jの場合)、第二方針(i=jの場合)にてカウントする遷移数N(Ci→Cj)は式(2)に示されている。また、式(1)における遷移確率P(Ci→Cj)の具体的な値は、遷移元クラスタCiの各々において、全ての遷移先クラスタCjを表すインデクスj(j=iの場合も含む)につき総和した式(3)の規格化条件を満たすように計算すればよい。式(2)にて右辺の全体を覆っている"||"は数学記号として周知のように集合の元の数を表す記号であって、例えば、|A|で集合Aに属する元の数を表す記号である。また式(2)にて、数学表記として周知のように、{x|xが満たす条件}で当該条件を満たすxの集合を表す。
【００６４】
【数２】

$$P(Ci \to Cj) \propto N(Ci \to Cj) \qquad \cdots(1)$$

$$N(Ci \to Cj) = \left|\left\{(X,n)\middle|D(X,n) \in Ci \text{ かつ } D(X,n+1) \in Cj\right\}\right| \qquad \cdots(2)$$

$$\sum_j P(Ci \to Cj) = 1 \qquad \cdots(3)$$

【００６５】
図８に、クラスタリング部51におけるクラスタリング結果及び対応する遷移確率の算出の例を[1]～[5]と分けて示す。[1]は、当該例におけるクラスタリング対象となった全データの例であり、Aさんに関して40歳～43歳の4個のデータD(A,40)～D(A,43)と、Hさんに関して43歳～46歳の4個のデータD(H,43)～D(H,46)と、の全8個のデータがクラスタリング対象であるものとする。（なお、予測モデルを構築するために一般にはもっと多数のデータを用いるが、ここでは算出例の説明のため、全データを8個としている。）
【００６６】
また、図８では「D(A,40)」等によりK番目の中間層におけるデータ「$D(A,40)_{[k]}$」を意味するものとする。（図８ではK依存の旨の表記は煩雑となるため省略しているものとする。）
【００６７】
[2]は、[1]の全データのクラスタリング結果であり、2つのクラスタC1,C2に分けられ、所属データがC1={D(A,40), D(A,41), D(A,42), D(H,45), D(H,46)}及びC2={D(A,43), D(H,43), D(H,44)}となっている。なお、[2]でデータ間に描いている矢印は、同一対象者デ

ータであって隣接年代n,n+1となっているデータ間に、年齢の進む方向に描いたものであり、次の[3]における遷移数のカウントの対象となる箇所に該当する。

【００６８】

[3]では、[2]のクラスタリング結果のクラスタC1,C2において一連の遷移数N(Ci→Cj)をカウントした結果が示されている。[4]は当該[3]のカウント結果を、遷移元クラスタを行要素とし、遷移先クラスタを列要素として行列形式に並べたものである。[5]には、当該[4]の結果を上記の式(3)を満たすように規格化した結果として、[2]のクラスタリング結果に対応する遷移行列（クラスタ間遷移確率を要素とする行列）が示されている。

【００６９】

[5]ではすなわち、遷移元がクラスタC1である一連の遷移のカウントN(C1→C1)=3及びN(C1→C2)=1を規格化することで遷移確率P(C1→C1)=0.75及びP(C1→C2)=0.25を求め、同様に、遷移元がクラスタC2である一連の遷移のカウントN(C2→C1)=1及びN(C2→C2)=1を規格化することで遷移確率P(C2→C1)=0.5及びP(C2→C2)=0.5を求めている。

【００７０】

以上のようにクラスタリング結果の各クラスタとそのクラスタ間の遷移確率を与えたものは、一実施形態では健康状態の推移モデルとして利用可能であり、クラスタリング部51においてこのような予測モデルを構築しておく。最適クラスタ数決定部53では当該構築された予測モデルのうち、最適クラスタ数$m_{[最小]}$におけるものを予測部6やユーザへと出力すればよい。

【００７１】

図９に、健康状態の推移モデルの例を示す。ここで、クラスタリング結果はC1,C2,C3の3クラスタであった場合を例としている。予測対象者につき、n歳の時点（現在）ではクラスタC1の健康状態であり、n+1歳の時点ではクラスタC2の健康状態であり、n+2歳の時点ではクラスタC2の健康状態であり、n+3歳の時点ではクラスタC3の健康状態である、という形で、当該クラスタリング結果をモデルとして用いた際の対象者の健康状態の推移が予測される。当該予測を実施するための具体的な計算等については、予測部6の説明において後述する。

【００７２】

AIC計算部52では、K番目の中間層のデータ$D(X,n)_{[K]}$よりクラスタ数m毎に得られたクラスタリング結果$CL(K)_{[m]}$よりクロス集計表を作成して、そのAIC値（赤池情報量基準の値）（AIC(K,m)とする）を計算する。

【００７３】

図１０に、当該作成するクロス集計表を示す。図１０では、クラスタリング結果$CL(K)_{[m]}$の各クラスタ1～mについて、属する対象者（(X,n)で指定される対象者Xのn歳時点の状態に対応）のうち、翌年に特定の疾病（糖尿病など）を発症した人数と、発症しなかった人数と、をカウントすることで作成されるクロス集計表が示されている。

【００７４】

なお、「特定疾病を発症したか否か」については、文書化部1が出力する全データD(X,n)に対して、事前に紐付いた情報として与えておくものとする。「特定疾病を発症したか否か」の他にも「当該年度における医療費が高いか否か」等の基準でクロス集計表を作成してもよい。後段側の予測部6において実施する予測に関連した事項で、健康状態の良い又は悪いに関する事項を基準とすることができる。

【００７５】

AIC計算部52では図１０のように作成されたクロス集計表に記載の各度数を入力として、AICの値を計算することができる。具体的な計算方法は後述する。

【００７６】

前述のようにAIC値が最小のものが最適クラスタ数決定部53において最適なクラスタリング結果として決定されるが、クロス集計表を図１０のように作成することで、「特定疾病を発症したか否か」等についてのモデル化が適切に行われているクラスタリング結果が最適な結果として選別されることとなる。

【００７７】
予測部6では、モデル構築部5の出力した予測モデルPM(K)を用いて、健康状態に関連する予測を行う。前述のように、図２、図３の実施形態の予測が可能であるので、以下それぞれ説明する。
【００７８】
図２の実施形態では、まず、分類評価部61が予測モデルPM(K)のクラスタリング結果の各クラスタを危険者クラスタ群と非危険者クラスタ群とに分ける。次に、危険判定部62が、ユーザより受け取った予測対象のデータにつき、PM(K)のクラスタリング結果におけるいずれのクラスタに所属するかを特定することで、当該所属クラスタが危険者クラスタ群に属するものであれば当該予測対象者は健康状態が「危険」にあるものと判定し、当該所属クラスタが非危険者クラスタ群に属するものであれば当該予測対象者は健康状態が「非危険」にあるものと判定することで、２値的な健康状態に関する予測を行う。
【００７９】
なお、「危険／非危険」との２値的な結果のみではなく、所属クラスタも出力することで、危険判定部62ではより詳細な健康状態に関する予測結果を出力することもできる。各部61,62の詳細は以下の通りである。
【００８０】
分類評価部61では、まず、予測モデルPM(K)のクラスタリング結果の各クラスタに関し、図１０のクロス集計表を作成したのと同様の基準で、「特定疾病を発症した」等の「発症者」の割合を求め、発症率の高い順番に並べる。さらに、図１１に例を示すように、[1]のように発症率の高い順に並べた結果がクラスタ1,2,…,mであったとする場合に、上位のK番目までのクラスタを仮の「危険クラスタ群」とし、K+1番目以降を同様に仮の「非危険クラスタ群」とする。
【００８１】
そして、図１１の[2]に示すように、[1]にて区別された「危険クラスタ群」と「非危険クラスタ群」とに関して、属するデータにつき図１０のクロス集計表と同様の基準の「発症者」及び「非発症者数」をカウントしたクロス集計表を作成する。
【００８２】
図１１の[2]のクロス集計表は危険判定基準とした「上位K」のKの値ごとに作成することができるので、それぞれのKの値につき分類評価部61ではAIC値を計算し、最小値となった際の$K=K_{[最小]}$を、実際に危険クラスタ群と非危険クラスタ群とを分けるのに適した結果として、危険判定部62へと出力する。
【００８３】
危険判定部62では、ユーザより予測対象として入力されるデータに関して、予測モデルPM(K)におけるいずれのクラスタに属するかを判断し、当該所属クラスタが危険クラスタ群と非危険クラスタ群のいずれであるかによって、前述のように予測を行うことができる。
【００８４】
なお、いずれのクラスタに属するかの判断は、予測対象データが既存データである場合と新規データである場合との各場合において、次のように行えばよい。まず、予測対象データが、モデルPM(K)の構築用データとして、文書化部1にて入出力された全データ内に含まれているいずれか１つのデータ（既存データ）である場合、当該データはクラスタリング部51でクラスタリングにより予測モデルPM(K)を構築した際のデータに含まれる（すなわちクラスタリング結果内に既に存在している）こととなるので、所属クラスタは自明である。一方、予測対象データが、モデルPM(K)の構築用データとして文書化部1が入出力したデータとは別のデータ（新規データ）である場合、次の第一処理及び第二処理を行うことで所属クラスタを判断すればよい。
【００８５】
第一処理では、当該新規の予測対象データ（文書化部1に対する入力データと同様の、特定対象者Xの特定年代nにおけるレセプト情報や健診情報として与えられている）に対し

て、文書化部１、第一正規化部2、データ表現学習部3、第二正規化部4までの処理を（予測モデルPM(K)の構築処理とは別途の処理として）行うことにより、そのバグオブワード形式を得る。すなわち、予測モデルPM(K)を既存データを用いてモデル構築部5において構築した際の入力となった、前述の第K中間層での表現データ$D(X,n)_{[K]}$を第二正規化部4で正規化した形式と同様の形式として、バグオブワード形式での予測対象データを得る。なおここで、データ表現学習部3の適用においては、ネットワーク重み最適化部32が（当該新規データ以外の既存データにより）既に計算済みである重みを用いて、中間層出力取得部33より第K中間層での表現データを得るようにすればよい。すなわち、既に構築済みの深層学習ネットワーク上において、第K中間層の表現データを得るようにすればよい。

【００８６】

第二処理では、第K中間層におけるバグオブワード形式で与えられた予測対象データを、予測モデルPM(K)におけるトピック比率の形に変換することにより、いずれのクラスタに属するかを判断する。ここで、バグオブワード形式の第K中間層での予測対象データ（予測モデルPM(K)における分解結果「D=θ×Φ」の行列「D」の行ベクトルに相当）に、予測モデルPM(K)における分解結果「θ×Φ」のうちの「Φ」行列の逆行列（ムーアペンローズの一般逆行列）を乗ずることにより、予測モデルPM(K)における予測対象データのトピック比率を求めることができる。

【００８７】

なお、第K中間層での表現データを正規化したものとしての、バグオブワード形式での予測対象データの準備は、予測対象データが新規のものである場合であっても、以上のように医療データ解析装置10において自動処理として実施することができるが、ユーザ側で当該準備を行うようにしてもよい。すなわち、ユーザ側のマニュアル作業等で別途、当該バグオブワード形式のデータを用意しておいたうえで、予測部6において当該用意されたバグオブワード形式での予測対象データに対して予測を実施するようにしてもよい。

【００８８】

なお、上記の結果は予測モデルPM(K)のK（中間層の深さを表すKであり、図１１の上位Kではない）ごとに求まるので、いずれのKの値の結果が最適なものかを判断するため、分類評価部61ではさらに、図１２のような集計表を作成し、当該集計表より統計分野において周知のF値を求め、F値が最小となるKが最適な結果であると判断してもよい。

【００８９】

図１２の集計表はその要素n_{11}等を記載しているように、図１１の[2]のAIC計算におけるものと同様である。図１１の[2]の要素n_{11}をそのまま図１２に記載の通りの値として採用することで、F値を計算することができる。

【００９０】

すなわち、図１２の集計表より精度Precision=TP/(TP+FP)を求め、再現率Recall=TP/(TP+FN)を求め、これらの調和平均としてF値=(2*Precision*Recall)/(Recall+Precision)を求めることができる。

【００９１】

図３の実施形態では、予測部6は前述の図９の健康状態の遷移の形で、予測を行うことができる。このため、遷移予測部65は以下のような計算を行えばよい。遷移予測部65に対し、ユーザ指示としては、次の第一指示～第三指示を与える。

【００９２】

まず、第一指示として、予測したい対象者（実際の対象者でも架空のものでもよい）の特定年代における健康データを入力として与える。当該入力は、文書化部1の出力と同様の単語頻度ベクトルの形式で与える。また、第二指示として、予測モデルPM(K)のうちのいずれを予測モデルとして用いるかを指示する。さらに、第三指示として、当該予測対象者に当該予測モデルを適用することで、その何年先の健康状態を予測するか、という指示を与える。以上の指示を与えることで、遷移予測部65では当該予測対象者に当該予測モデルを適用することで推定される、当該指定した年数だけ将来における健康状態を出力する

ことができる。
【００９３】
従って、上記の第一指示～第三指示を入力として受け取った遷移予測部65では、対象者データ（n歳時点でのデータとする）が当該n歳時点でいずれのクラスタに属するかと、n歳以降どのように所属クラスタを遷移するか、ということを計算する。当該計算は、前者（n歳での所属クラスタの決定）及び後者（n歳以降の所属クラスタの遷移）に関してそれぞれ以下のように実施することができる。
【００９４】
前者（n歳での所属クラスタの決定）に関しては、危険判定部62の説明におけるのと同様の手法で判定を行うことができる。
【００９５】
また、後者（n歳以降の所属クラスタの遷移）に関しては、前者で得られたn歳時点での所属クラスタを経路の始点に設定したうえで、図９に示すような状態遷移系列に対して、周知のビタビアルゴリズムを適用し、最大確率を与える経路として、クラスタ間遷移を計算することができる。ここで、状態間の遷移確率には、クラスタリング部51で求めたクラスタ間の遷移確率をそのまま利用すればよい。
【００９６】
なお、上記のようにビタビアルゴリズムで計算した結果、予測結果がどのクラスタに属する確率も同じ（閾値判定で同じ）ということになる場合もある。この場合、予測モデルが機能していない。そこで、期間設定部66では、遷移予測部65に対して、Kの値がより大きな予測モデルPM(K)を用いた予測を代わりに行うように指示し、ユーザ指定に対して、予測モデルが機能している（結果が等確率とはならない）ような予測結果が得られるまで、遷移予測部65に計算を継続させるようにすることができる。
【００９７】
期間設定部66の処理は、次のような性質に基づく。すなわち、一般的に抽象度が高いデータ（Kの大きい側のデータ）ほど、得られる最適クラスタ数は減少する。クラスタ数が減少すれば、より長期の予測が可能となる。
【００９８】
以下、本発明の効果の実例を紹介する。
【００９９】
■データについて
・健診データについては特定健診の項目：体重、身長、BMI、HbA1c等を用い、データを0～1に正規化して入力データとする
・問診データについては、基本的には「はい」「いいえ」等で表される2値～4値程度の離散データであり、これらを0,1で表現した。（2値の場合、0,1、4値の場合、0,0.25,0.5,0.75,1）また用いたデータは2009～2012年のデータとして、1年毎に分割して別々の入力データとする。
⇒これらを合わせると37次元データとなる。
・さらに実験では、DLの中間層を35、34・・・と1次元ずつ減らしていってデータ表現を抽象化した。抽象化後、得られたデータを最大値12で正規化し、LDAへの入力とする。
・また比較のために一般的なLDAのみを用いた場合を実験した。この場合には、全てのデータの最大値を12として、バグオブワーズ表現とする。
・2009年時点で生活習慣病を発症していない人が、2010・2011年時点で生活習慣病を発症する確率（発症率）を全てのクラスタで計算し、クラスタ数の最適化を行う。
・また上記発症率について各クラスタの精度・再現率を計算してF値を割り出し、評価を行う。（このとき、発症率が高いクラスタを順番に並べ、危険クラスタとみなす分割位置を変えながら、最適分割位置を決定する。）
【０１００】
■結果
・DLを用いた場合：

データの次元数を30まで抽象化した場合について、上記のF値を計算したところ、35まで抽象化したとき、最適クラスタ数が21となり、最もF値が改善された。平均よりも発症率が高いクラスタを集めて、ハイリスク者とみなした場合のF値は0.3。
・一般的なLDAを用いた場合：
最もF値が高い最適クラスタ数は8となり、F値は0.29。
同様の実験を健診データのみで行うと、DLを用いた場合には0.27、LDAのみの場合には0.28であり、精度向上は見込めていない。これは健診データが単一のデータ系列であるために、抽象化の必要性がなかったためと思われる。

【０１０１】

図１３に、上記結果のDL及びLDAの場合をそれぞれ[1],[2]として示す。なお、図１３では灰色で描かれているのが、平均よりも発症率が高いクラスタのデータである。

【０１０２】

以下、本発明における補足的事項を説明する。

【０１０３】

（１）AIC計算部52による図１０のクロス集計表を用いた、また、分類評価部61による図１１の[2]のクロス集計表を用いた、AIC値の算出について

【０１０４】

図１４は、図１０あるいは図１１のクロス集計表を一般化した表である。すなわち、図１０の集計数n_{ij}は図１０、図１１等と共通のものを一般の場合として示しており、何らかの基準に該当するか否かを縦軸（行要素）として、クラスタ等の分類結果を横軸（列要素）として、構成されている。

【０１０５】

図１４に示すように、当該クロス集計表における集計数n_{ij}により、ただちに周辺度数ki(i=1,2,…,m)、h,N-h等を計算することができ、これらの値を用いて以下のようにAIC値を計算することがきる。

【０１０６】

当該AIC値は、次のいずれかの手法の値として求める。第一手法では、当該クロス集計表に対して従属モデルを適用することにより、以下の[式1]のような従属モデルのAIC値AIC(DM)[ここでDMはDependent Modelの略である]として求める。第二手法では、さらに、当該クロス集計表に対して独立モデルを適用して、以下の[式2]のような独立モデルのAIC値AIC(IM)[ここでIMはIndependent Modelの略である]を求めたうえで、[式3]のように、従属モデルのAIC値から独立モデルのAIC値を引いた差の値として、求める。

【０１０７】

【数３】

$$\mathrm{AIC(DM)} = -2 \times \mathrm{MLL(DM)} + 2 \times (2\mathrm{m} - 1) \quad [式\ 1]$$

ここで、$\mathrm{MLL(DM)} = n_{11} \cdot \log(n_{11}) + n_{12} \cdot \log(n_{12}) + \cdots - \mathrm{N} \cdot \log(\mathrm{N})$ [式 1－2]

$$\mathrm{AIC(IM)} = -2 \times \mathrm{MLL(IM)} + 2 \times \mathrm{m} \quad [式\ 2]$$

ここで、MLL(IM)

$$= \mathrm{h} \cdot \log\left(\frac{h}{N}\right) + (\mathrm{N} - \mathrm{h})\log(1 - \frac{h}{N}) + k_1 \cdot \log\left(\frac{k_1}{N}\right) + \cdots + (\mathrm{N} - k_1 - k_2 - \cdots) \cdot \log(1 - \frac{k_1 + k_2 + \cdots}{N})$$

$$= \mathrm{h} \cdot \log(\mathrm{h}) + (\mathrm{N} - \mathrm{h}) \cdot \log(\mathrm{N} - \mathrm{h}) + k_1 \cdot \log(k_1) + \cdots + \left(\mathrm{N} - k_1 - k_2 - \cdots\right) \cdot \log\left(\mathrm{N} - k_1 - k_2 - \cdots\right) - 2 \cdot \mathrm{N} \cdot \log(\mathrm{N}) \quad [式\ 2-2]$$

$$\mathrm{AIC(DM)} - \mathrm{AIC(IM)} = -2 \times \mathrm{MLL(DM)} + 2 \times \mathrm{MLL(IM)} + (2\mathrm{m} - 4) \quad [式\ 3]$$

【０１０８】

なお、[式1]等においてMLL(DM)は、従属モデルにおける最大対数尤度であって、[式1-2]のような値として求めることができる。また、[式2]等において、MLL(IM)は、独立モデルにおける最大対数尤度であって、[式2-2]のような値として求めることができる。なお、上記の各式における文字は、図１２のクロス集計表において説明した通りであり、以降説明する各式においても同様である。

【０１０９】

以下、従属モデルにおける最大対数尤度MLL(DM)と、独立モデルにおける最大対数尤度MLL(IM)と、がそれぞれ、上記の[式1-2]及び[式2-2]のように算出されることと、当該算出されたそれぞれの最大対数尤度を用いて、従属モデルにおけるAIC値が[式1]のように算出され、また、独立モデルにおけるAIC値が[式2]のように算出されることを説明する。

【０１１０】

図１５は、[式1]及び[式1-2]として示した従属モデルにおける算出を説明するための、図１４のクロス集計表に対応する従属モデルにおける確率の表である。当該表に示されている確率により、以下のように算出がなされる。

【０１１１】

まず、従属モデルの確率変数は以下の通りである。

【０１１２】

【数４】

$$\Pr(n_{11}, n_{12}, \cdots n_{1m}, n_{21}, n_{22}, \cdots n_{2m}) = \left(\frac{n_{11}}{N}\right)^{n_{11}} \cdot \left(\frac{n_{12}}{N}\right)^{n_{12}} \cdots \left(\frac{n_{2m}}{N}\right)^{n_{2m}}$$

$$= {p_{11}}^{n_{11}} \cdot {p_{12}}^{n_{12}} \cdots {p_{2\mathrm{m}}}^{n_{2m}}$$

【０１１３】

一方、図１５に示された2m個の全てが自由に動かせるわけではなく、以下の制約がある。

【０１１４】

【数５】

$$p_{2m} = 1 - (p_{11} + p_{12} + \cdots)$$

【０１１５】

従って、従属モデルの自由度は2m-1であり、AICの定義（AIC＝-2×MLL+2×自由度）より、[式1]の2*(2m-1)の項が得られる。さらに、上記確率変数より対数尤度LLを計算すると、以下のようになる。

【０１１６】

【数６】

$$\mathrm{LL} = \log\left({p_{11}}^{n_{11}} \cdot {p_{12}}^{n_{12}} \cdots\right) = n_{11} \cdot \log(p_{11}) + n_{12} \cdot \log(p_{12}) \cdots + n_{2m} \cdot \log(p_{2m})$$

【０１１７】

上記対数尤度LLを最大にするときの条件は以下である。

【０１１８】

【数７】

$$\frac{\delta LL}{\delta p_{11}} = \frac{\delta LL}{\delta p_{12}} = \cdots = 0$$

【０１１９】

上記最大とする条件より、以下が得られる。

【０１２０】

【数８】

$$\frac{\delta LL}{\delta p_{11}} = \frac{n_{11}}{p_{11}} - \frac{n_{2m}}{p_{2m}} = 0$$

【０１２１】

上記と同様にして、さらに

【０１２２】

【数９】

$$\frac{n_{12}}{p_{12}} - \frac{n_{2m}}{p_{2m}} = 0$$

【０１２３】

等が得られる。そこで、

【０１２４】

【数１０】

$$\frac{n_{2m}}{p_{2m}} = \mathrm{r}$$

【０１２５】

とすると、

【０１２６】

【数１１】

$$\mathrm{n}_{11} = p_{11} \cdot r$$

$$\mathrm{n}_{12} = p_{12} \cdot r$$

【０１２７】

等となるので、それぞれを足すと、

【０１２８】

【数１２】

$$n_{11} + \mathrm{n}_{12} + \cdots = (p_{11} + p_{12} + \cdots)\mathrm{r} = \mathrm{r} = \mathrm{N}$$

【０１２９】

となるから、以下の場合が最尤推定となる。

【０１３０】

【数１３】

$$p_{11} = \frac{n_{11}}{N}, p_{12} = \frac{n_{12}}{N}, \cdots$$

【０１３１】

従って、上記の値をLLに代入することで、その最大値として前述の[式1-2]が得られる。

【０１３２】

図１６は、[式2]及び[式2-2]として示した独立モデルにおける算出を説明するための、図１４のクロス集計表に対応する従属モデルにおける確率の表である。当該表に示されている確率により、以下のように算出がなされる。

【０１３３】

まず、図１４の周辺度数k_mと、対応する図１６の周辺確率q_mと、において、以下のような制約がある。

【０１３４】

【数１４】

$$k_m = n_{1m} + n_{2m} = N - k_1 - k_2 - \cdots$$

$$q_m = 1 - q_1 - q_2 - \cdots$$

【０１３５】

従って、自由に動かせるのはq_1～q_{m-1}とpとであるから、パラメータの自由度は(m-1)+1=mであって、AIC算出の定義より、[式2]の2×mの項が得られる。また、独立モデルの確率変数は以下の通りとなる。

【０１３６】

【数１５】

$$\Pr(n_{11}, n_{12}, \cdots n_{1m}, n_{21}, n_{22}, \cdots n_{2m}) = \mathrm{p}^{h} \cdot q_1^{k_1} \cdot q_2^{k_2} \cdots (1 - q_1 - q_2 - \cdots) \cdot (\mathrm{N} - k_1 - k_2 - \cdots)$$

【０１３７】

従って、その対数尤度LLは以下の通りとなる。

【０１３８】

【数１６】

$$\mathrm{LL} = \mathrm{h} \cdot \log(\mathrm{p}) + (\mathrm{N} - \mathrm{h}) \cdot \log(1 - \mathrm{p}) + k_1 \cdot \log(q_1) + k_2 \cdot \log(q_2) + \cdots + (\mathrm{N} - k_1 - k_2 - \cdots) \cdot \log(1 - q_1 - q_2 - \cdots)$$

【０１３９】

対数尤度の最大値を与える条件を求めるべく、これをp、q_1・・・で偏微分してゼロに等しいとすることにより、以下等の一連の計算ができる。

【０１４０】

【数１７】

$$\frac{\delta LL}{\delta p} = \frac{\delta LL}{\delta q_1} = \cdots = 0$$

$$\frac{h}{p} - \left(\frac{N - h}{1 - p}\right) = 0$$

$$\frac{k_1}{q_1} - \left(\frac{N - k_1 - k_2 - \cdots}{1 - q_1 - q_2 - \cdots}\right) = 0$$

$$\frac{k_2}{q_2} - \left(\frac{N - k_1 - k_2 - \cdots}{1 - q_1 - q_2 - \cdots}\right) = 0$$

$$\cdots$$

【０１４１】

従って、

【０１４２】

【数１８】

$$\mathrm{p} = \frac{h}{N}$$

【０１４３】

となり、また、

【０１４４】

【数１９】

$$\frac{k_1}{q_1} = \cdots = \mathrm{r} = \left(\frac{N - k_1 - k_2 - \cdots}{1 - q_1 - q_2 - \cdots}\right)$$

【０１４５】

とすると、

【０１４６】

【数２０】

$$\mathrm{k}_1 = \mathrm{q}_1 \cdot r$$

$$\mathrm{k}_2 = \mathrm{q}_2 \cdot r$$

$$\cdots$$

【０１４７】

等となるので、それぞれ足して、
【０１４８】
【数２１】

$$k_1 + k_2 + \cdots = \mathrm{r}\left(q_1 + q_2 + \cdots\right) = \mathrm{N}$$

【０１４９】
となり、
【０１５０】
【数２２】

$$\mathrm{r} = \mathrm{N}$$

【０１５１】
となるから、最大尤度は
【０１５２】
【数２３】

$$q_1 = \frac{k_1}{N}$$

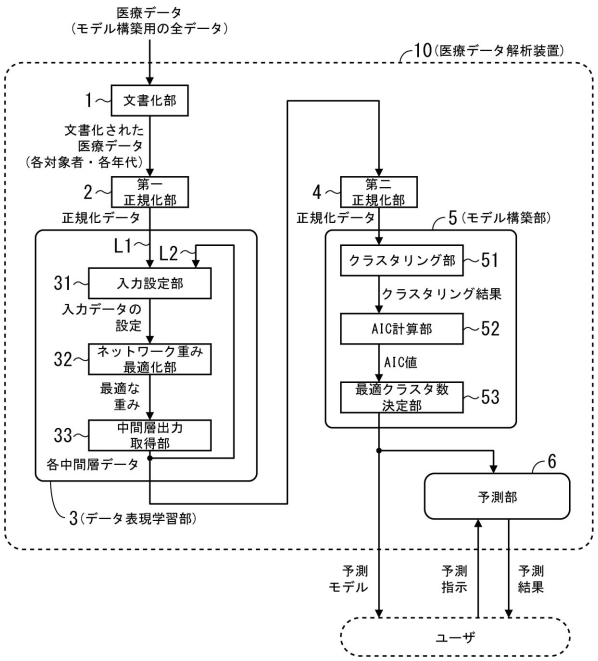
$$q_2 = \frac{k_2}{N}$$

$$\cdots$$

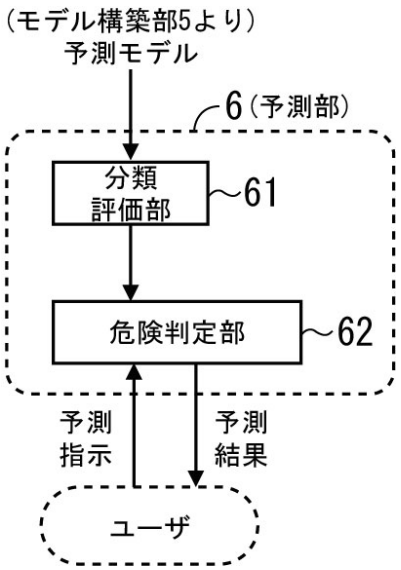
【０１５３】
等において得られることとなる。従って、上記の値をLLに代入することで、最大値としての[式2-2]が得られる。
【０１５４】
（２）文書化部1では、各対象者Xの各年代n（年齢n）における文書化された医療データD(X, n)を生成するものとし、当該年代nは1年毎に与えられているものとして以降の説明を行ったが、1年に限らず、任意の長さの所定期間（２年あるいは半年など）ごとの年代nで区切ってデータD(X,n)を生成してもよい。この場合、図９で説明したような健康推移モデルの推移のステップ幅も、当該任意の長さの所定期間となる。例えば、2年毎のデータD(X,n)を利用する場合、健康推移モデルは2年毎の状態を与えるものとなる。
【０１５５】
（３）文書化部1では、入力される健診データその他の医療データを、各対象者Xの各年代nにおける健康状態に対応するバグオブワードとしてのデータD(X,n)に変換するものとして説明したが、入力されるデータが予め当該バグオブワードの形式に変換されている場合、文書化部1は省略されてもよい。
【０１５６】
（４）図１０のクロス集計表は、「翌年」に特定疾病を発症するか否かという基準で作成する場合を例として説明したが、この場合、予測部6でも当該基準に基づく予測を実施するようにすることが好ましい。（その理由は、AIC値に基づき当該基準に関して適した分類結果が予測モデルとして選択されているからである。）すなわち、図２の実施形態であれば予測対象データが翌年に特定疾病を発症するか否かに関して危険群に属するかそうでないかということを予測し、図３の実施形態であれば翌年の健康状態を予測するようにすることが好ましい。
【０１５７】
従って、「翌年」に限らず任意の期間の経過後、例えば「ｎ年後」に特定疾病を発症するか否かという基準で図１０のクロス集計表を作成するようにしてもよい。この場合、予測部6を図３の実施形態として実現する場合、１年ごとの状態遷移予測をn回繰り返すことでn年後の健康状態を予測するようにしてもよいし、１回の状態遷移がn年の経過に対応するものとして、１回の状態遷移でn年後の健康状態を予測するようにしてもよい。

【０１５８】
（５）本発明は、コンピュータを医療データ解析装置10の各部1～6の全て又はその任意の一部分として機能させるプログラムとしても提供可能である。当該コンピュータには、CPU(中央演算装置)、メモリ及び各種I/Fといった周知のハードウェア構成のものを採用することができ、CPUが医療データ解析装置10の各部の機能に対応する命令を実行することとなる。
【符号の説明】
【０１５９】
10…医療データ解析装置、1…文書化部、2…第一正規化部、3…データ表現学習部、4…第二正規化部、5…モデル構築部、6…予測部

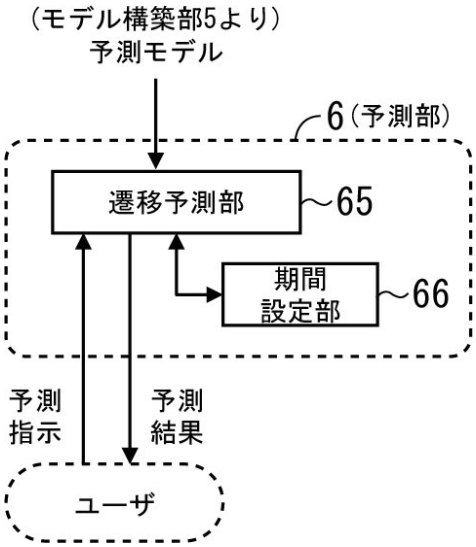
【図１】



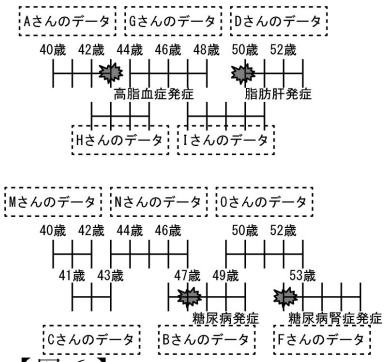
【図２】



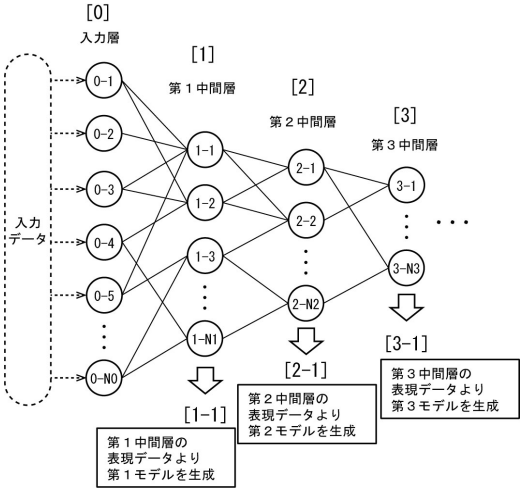
【図３】



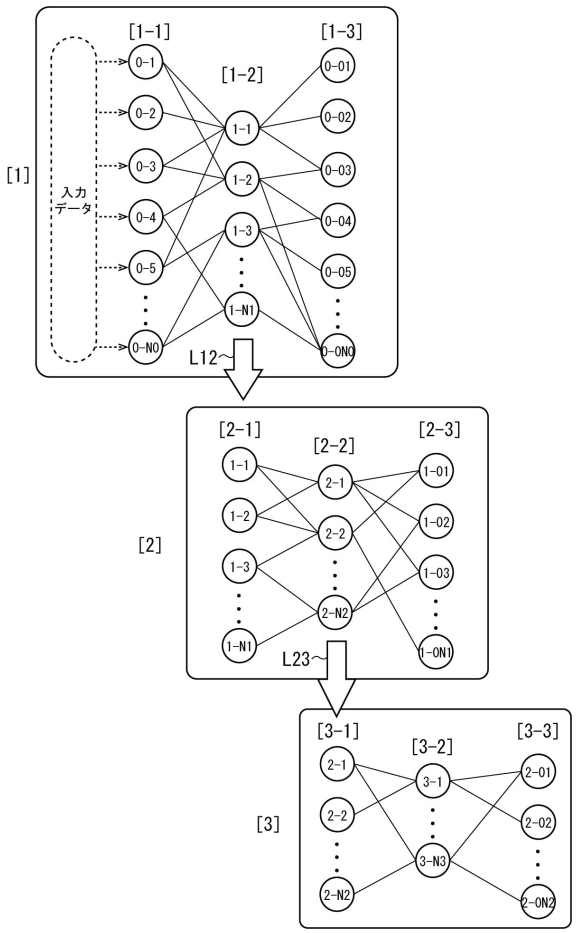
【図４】



【図５】



【図６】

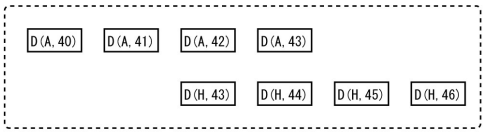


【図７】

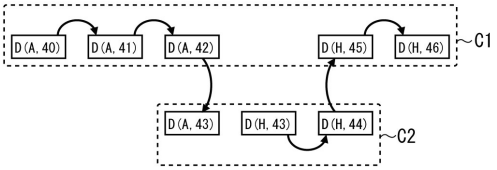
$$\begin{pmatrix} \text{文書:u} & \overset{\text{単語:i}}{D} \end{pmatrix} = \begin{pmatrix} \text{文書:u} & \overset{\text{トピック:k}}{\theta} \end{pmatrix} \times \begin{pmatrix} \text{トピック:k} & \overset{\text{単語:i}}{\Phi} \end{pmatrix}$$

【図８】

[1]



[2]



[3]

遷移数	該当する遷移の列挙
N(C1→C1)=3	D(A, 40)→D(A, 41)，D(A, 41)→D(A, 42)，D(H, 45)→D(H, 46)
N(C1→C2)=1	D(A, 42)→D(A, 43)
N(C2→C2)=1	D(H, 43)→D(H, 44)
N(C2→C1)=1	D(H, 44)→D(H, 45)

[4]

遷移先クラスタ

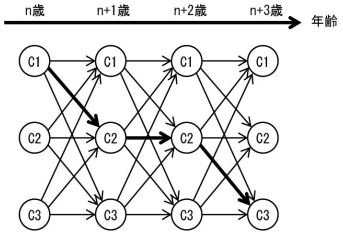
遷移元クラスタ	C1	C2
C1	3	1
C2	1	1

[5]

遷移先クラスタ

遷移元クラスタ	C1	C2
C1	0.75	0.25
C2	0.5	0.5

【図９】



【図１０】

	クラスタ1	クラスタ2	…	クラスタm
翌年に特定の疾病を発症した人数	n_{11}	n_{12}	…	n_{1m}
非発症の人数（健康者の人数）	n_{21}	n_{22}	…	n_{2m}

【図１１】

[1]

	クラスタ1	クラスタ2	…	クラスタK	クラスタK+1	…	クラスタm
発症率	0.8	0.7	…	0.4	0.3	…	0.1

危険クラスタ群 クラスタ1～K　　非危険クラスタ群 クラスタK+1～m

[2]

	危険クラスタ群（クラスタ1～K）	非危険クラスタ群（クラスタK+1～m）
発症者数	n_{11}	n_{12}
非発症者数	n_{21}	n_{22}

【図１２】

	予測の結果が危険クラスタ群に所属する人数（Positive）	予想の結果が非危険クラスタ群に所属する人数（Negative）
実際に発症した人数（True）	TP(=n_{11})	TN(=n_{12})
実際には発症しなかった人数（False）	FP(=n_{21})	FN(=n_{22})

【図１３】

[1]

既存LDA	発症者	人数	各クラスタの精度
クラスタ0	74	529	0.139886578
クラスタ1	183	1138	0.160808436
クラスタ2	345	1590	0.216981132
クラスタ3	0	0	0
クラスタ4	78	504	0.154761905
クラスタ5	84	757	0.110964333
クラスタ6	23	271	0.084870849
クラスタ7	0	1	0
合計	787	4790	0.164300626

[2]

DL+LDA	発症者	人数	各クラスタの精度
クラスタ0	45	284	0.158450704
クラスタ1	34	193	0.176165803
クラスタ2	41	199	0.206030151
クラスタ3	34	196	0.173469388
クラスタ4	67	293	0.228668942
クラスタ5	35	270	0.12962963
クラスタ6	22	219	0.100456621
クラスタ7	26	124	0.209677419
クラスタ8	43	228	0.188596491
クラスタ9	50	237	0.210970464
クラスタ10	31	185	0.167567568
クラスタ11	54	207	0.260869565
クラスタ12	48	327	0.146788991
クラスタ13	21	158	0.132911392
クラスタ14	37	262	0.141221374
クラスタ15	40	293	0.136518771
クラスタ16	20	246	0.081300813
クラスタ17	45	272	0.165441176
クラスタ18	18	183	0.098360656
クラスタ19	36	254	0.141732283
クラスタ20	40	160	0.25
合計	787	4790	0.164300626

【図１４】

m個のクラスタに分けられたクラスタリング結果において
各クラスタの評価指標の「該当」「未該当」の度数を与えて得られる
2×mのクロス表

	クラスタ1	クラスタ2	…	クラスタm	周辺度数
「該当」度数	n_{11}	n_{12}	…	n_{1m}	h
「未該当」度数	n_{21}	n_{22}	…	n_{2m}	N-h
周辺度数	k_1	k_2	…	k_m	N

【図１５】

[従属モデルにおける確率の表]

	クラスタ1	クラスタ2	…	クラスタm	周辺確率
「該当」確率	p_{11}	p_{12}	…	p_{1m}	p
「未該当」確率	p_{21}	p_{22}	…	p_{2m}	1-p
周辺確率	q_1	q_2	…	q_m	1

【図１６】

[独立モデルにおける確率の表]

	クラスタ1	クラスタ2	…	クラスタm	周辺確率
「該当」確率	pq_1	pq_2	…	pq_m	p
「未該当」確率	$(1-p)q_1$	$(1-p)q_2$	…	$(1-p)q_m$	1-p
周辺確率	q_1	q_2	…	q_m	1

フロントページの続き

(72)発明者　松本　一則
埼玉県ふじみ野市大原二丁目１番１５号　株式会社ＫＤＤＩ研究所内

審査官　松尾　真人

(56)参考文献　特開２０１４－２２５１７６（ＪＰ，Ａ）
特開２０１５－０９０６８９（ＪＰ，Ａ）
特開２０１４－１７８８００（ＪＰ，Ａ）
遠藤　結城，移動手段判定のための表現学習を用いたＧＰＳ軌跡からの特徴抽出，Ｗｅｂとデータベースに関するフォーラム　情報処理学会シンポジウムシリーズ　Ｖｏｌ．２０１４　Ｎｏ．４　［ＣＤ－ＲＯＭ］　WebDB Forum 2014，一般社団法人情報処理学会，２０１４年１１月１２日，第２０１４巻，第４号
畠山　豊，問診データに対する潜在トピックモデルに基づく健診データ解析　Analysis of Health Checks Data Based on Latent Topic Model for Medical Interview，医療情報学　，一般社団法人日本医療情報学会／株式会社篠原出版新社，２０１３年１２月２６日，第３３巻，第５号，ｐ．２６７－２７７

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０
Ｇ０６Ｑ　５０／２２
Ｇ１６Ｈ　１０／００