個人健康管理を目的とした健康データマイニングシステム

竹内裕之† 児玉直樹† 橋口猛志‡ 林 同文‡

†高崎健康福祉大学 健康福祉学部 医療福祉情報学科 〒370-0033 群馬県高崎市中大類町 37-1
‡東京大学大学院 医学系研究科 健康医科学創造講座 〒113-8655 東京都文京区本郷 7-3-1

E-mail: †htakeuchi@takasaki-u.ac.jp, †kodama@takasaki-u.ac.jp, ‡hashiguchi-mi@umin.ac.jo, ‡hayashi-tky@umin.ac.jp

あらまし 個人の健康管理を目的とし、生活習慣と健康状態に関する時系列データを蓄積し生活習慣と健康状態の相関ルール解析を行う健康データマイニングシステムを開発した．健康に少し不安がある、あるいは健康増進をしたい一般健常者を対象としたシステムであり、家庭で取得した日常のデータを携帯電話からインターネット経由で解析を行うサーバに蓄積し、解析結果を携帯電話に返すサービスの実現を目指している．食事による摂取エネルギー、運動による消費エネルギー、睡眠時間、飲酒量、喫煙量など生活習慣データ項目を入力変数（独立変数）とし、血圧、体重、体脂肪率など健康データ項目を出力変数（ターゲット変数）として、決定木とアソシエーションルール解析によりルールを抽出した．

キーワード Web とインターネット、データマイニング、健康 DB

1. はじめに

1.1. 研究の背景

最近リモートの患者をケアする遠隔医療への Web テクノロジーの応用が世界的に注目を集めている[1-3]．急速に発展した Web テクノロジーは、さらに予防医学や健康管理の面でもその活用が期待されている．特に我が国では少子高齢化の進展が著しく、国民一人一人が健康で長生きして若年層の負担を軽減することが喫緊の課題となっており、病気の一次予防や健康増進のために日常の個人ベースの健康管理が必要であることが指摘されている[4]．これは、病気に関するリスクが遺伝的体質や生活環境により一人一人異なるためであり、正にテーラーメイドの健康維持・管理手法の開発が必要になっている．

1.2. 研究の概要

そこで、我々は携帯電話と Web テクノロジーを活用した個人健康管理システムを開発した[5,6]．このシステムは、携帯電話を端末として日常の生活習慣データと健康データをインターネット経由で時系列的にサーバコンピュータに蓄積する仕組である．蓄積されたデータはその統計を判りやすいグラフ表示で見ることができる他、個人の生活習慣と健康状態の間に何らかの規則性が見出せれば相関ルールとしてユーザの携帯電話に通知する．そして、これらの情報を参考にユーザが自分で自分の健康管理を行うことを期待している．本論文では、開発した個人健康管理システムにおいて、時系列的に蓄積されたデータから生活習慣と健康状態の間に規則性を見出す手法（健康データマイニング）について検討した結果を報告する．

2. 研究の方法

2.1. 個人健康管理システムの構成

システムの構成を図１に示す．Web クライアントを携帯電話に限定しているが典型的な MVC アーキテクチャの Web サイトである．筆者の大学（高崎健康福祉大学）の LAN とは独立に研究室に小規模の LAN を構成している．１台目のサーバ機はファイアウォールと DNS の役割を、２台目のサーバ機は Web サーバ、メールサーバおよび DB サーバの役割を、３台目のサーバ機はデータマイニングサーバの役割をそれぞれ持つ．１台のクライアント機は SPSS 社のデータマイニングツール「Clementine」を用い手動でデータ解析を行うための DB クライアントである．

生活習慣データや健康データといった個人情報を扱うため、Web クライアントである携帯電話と Web サーバ間の通信は暗号化された HTTPS プロトコルを用いている．またシステムの利用者は不特定多数なので通常のパスワード認証を用いている．

2.2. 健康データマイニングのコンセプト

健康データマイニングのコンセプトを図２に示す。ここでは、個人の現在の健康状態がなんらかの形で日常の生活習慣の影響を受けていると仮定する．そして、その関係は複数の項目が絡んだ複雑なもので、個人差も大きいものと考える．健康データマイニングの目的

は、日常の生活習慣データと健康データを個人毎に時系列的に蓄積し、その中から生活習慣と健康状態の間になんらかの規則性を見出し個人毎のルールとして抽出することである[7].

したがって、健康データマイニングでは、生活習慣データ項目を入力変数（独立変数）、健康データ項目を出力変数（ターゲット変数）として位置付け、「生活習慣データ $Y=y$ ならば健康データ $X=x$ の傾向がある」といったルールを個人毎に抽出する．即ちルールの前提部には生活習慣データ項目が、結論部には健康データ項目が含まれる．システムのユーザはこのような個人毎のルールを自己の健康管理や健康増進のために役立てることができる．

なおここではルールをシンプルにするために生活習慣データ項目間の関連性については考慮しなかった．

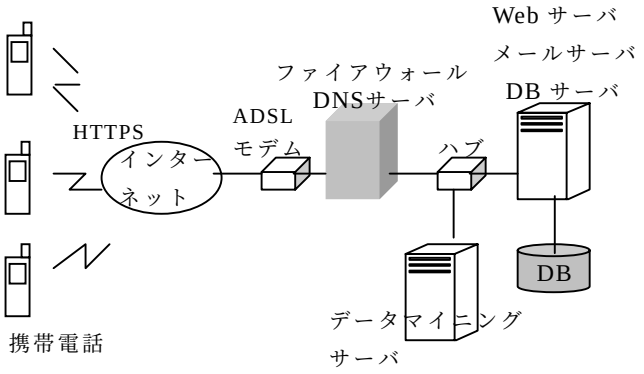

図１．個人健康管理システムの構成

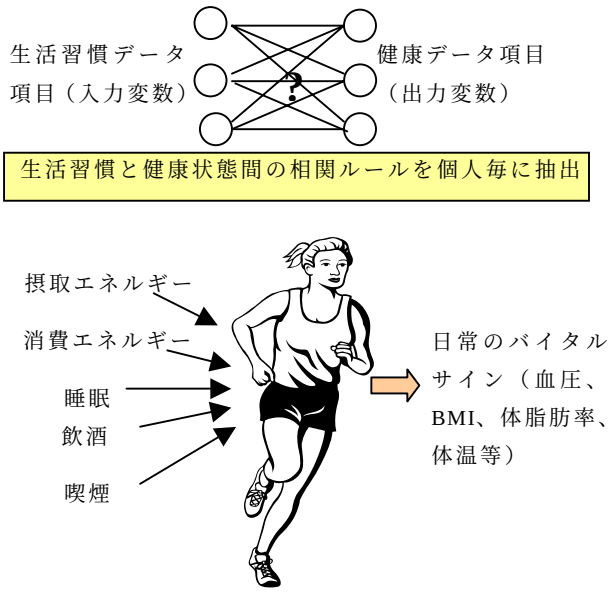

図２．健康データマイニングのコンセプト

2.3. 健康データマイニングプロセス

本研究での健康データマイニングにおいては、ルールの抽出過程が明瞭な決定木と、シンプルな if-then ルールを抽出するアソシエーションルール解析を試みた．前者は C5.0 アルゴリズム[8,9]を用い、後者はルールの自動抽出に適している ITRULE アルゴリズム[10]を用いた．

2.3.1. 入力変数の定義

決定木、アソシエーションルール解析ともに如何に有効な入力変数を定義するかがポイントとなる．本研究の最終目標は、インターネット上の不特定多数ユーザの大量の日常データを処理することなので、あまりにも多くの（有効でない）入力変数はいたずらにコンピュータの処理時間を消費するだけである．

そこで、関心のある健康データ項目（出力変数）と入力変数になりうる生活習慣データ項目の時系列データをもとに、相関を事前にチェックし入力変数（フィールド）を自動定義する手法を開発した．以下、図３を用い入力変数（フィールド）の自動定義のアルゴリズムについて説明する[7].

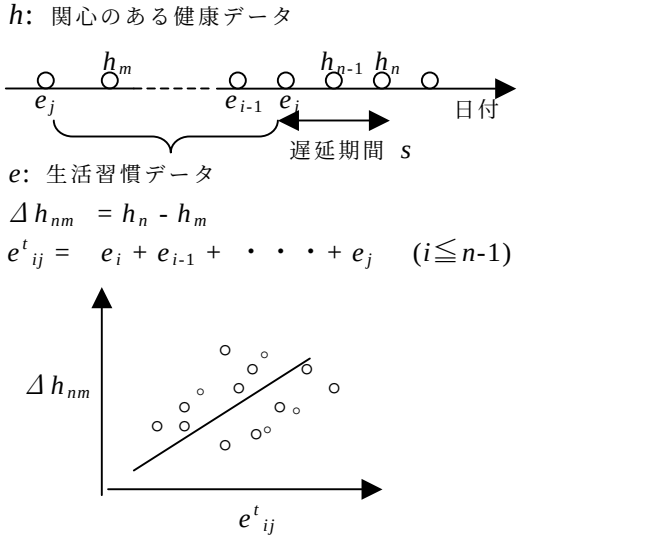

図３．入力フィールド自動定義のアルゴリズム

図３において、hは関心のある健康データ項目を、eは入力変数（フィールド）の候補となる生活習慣データ項目を表す．h_nはhのn日におけるデータ値、e_iはeのi日におけるデータ値、sは遅延期間である．ここで、

$$\Delta h_{nm} = h_n - h_m \quad (1)$$

なる量と、

$$e^t{}_{ij} = e_i + e_{i-1} + \cdots + e_j \quad (2)$$

なる量を定義する．遅延期間は $s = n - i$ で定義する．次に蓄積された時系列データについて、n-m, i-j, sをパラメータとして変化させながら(n-m=1~10, i-j=0~9, s=1~3)、Δh_{nm}と$e^t{}_{ij}$の間のピアソンの積率相関係数を計算する．ここで、n日より前の生活習慣データがn日の健康データに影響を与えると仮定する($i \leqq n$-1)．各(n-m, i-j, s)のセットにつきピアソンの積率相関係数を計算し、もし、１つ以上の相関係数がある閾値R_sより大きいものがあれば、そのeをhに対する入力変数として採用する．そして、実際の入力フィールドは相関係数が最大となる(n-m, i-j, s)のセット((n-m)$_{\max}$, (i-j)$_{\max}$, $s_{\max}$)をもとに自動定義する。例えば、(i-j)$_{\max}$=2, $s_{\max}$=2 であれば、eに関わる入力フィールドを

$$e_i + e_{i-1} + e_{i-2}\ (i=n\text{-}2) \quad (3)$$

と自動定義する．(i-j)$_{\max}$が大きいということは、長期間の生活習慣データの蓄積が現在の健康データに影響を与え、$s_{\max}$が大きいということは、生活習慣データが遅れをもって現在の健康データに影響を与えるということになる．

生活習慣データ項目eの値が数値でなくシンボル値の場合は、時系列データに基づくhとの相関係数は、シンボル値を適当に数値に変換して計算する．例えば、シンボル値が“多い”、“普通”、“少ない”、であれば、それぞれ 3, 2, 1 と変換する．例として、e_i=”多い“、e_{i-1}=”少ない“、であれば、$e_i + e_{i-1}$ = 3 + 1 =4 とする．ただし、入力変数として採用されルールマイニングを行うときには入力フィールドの値はシンボル値をそのまま用いる．

2.3.2. 出力変数の定義

健康データマイニングにおける出力変数（フィールド）は、関心ある健康データ項目である．健康データ項目hの値がシンボル値の場合は、Δh_{nm}を計算するときには、2.3.1（前項）で述べたシンボル値を持つ生活習慣データ項目と同じように数値に変換する.ただし、ルール生成プロセスでは出力変数（フィールド）値はシンボル値をそのまま用いる．

一方、健康データ項目 h の値が数値の場合は、全ての時系列データを“高い”、“中間”、“低い”というシンボル値を持つ３つのクラスに分類する．この分類において３つのクラスの境界値は、それぞれのクラスのデータ頻度が同程度になるように自動設定する．

2.3.3. ルール自動生成の手法

開発した個人健康管理システムでは、インターネット経由で不特定多数のユーザが日常のデータを蓄積し、それに対してタイミングよくデータ解析（健康データマイニング）を行い有効なルールを提示する必要がある．従って、データがある程度蓄積されると自動的に実行されることがサービスとして必須である.そこで、我々は決定木の他に、多くのデータセットの中から自動的にルールを抽出するのに適した手法として、情報理論に基づくアソシエーションルール解析手法であるITRULE アルゴリズム[10]を用いることにした．

ITRULE アルゴリズムは、

$$\text{If } Y=y, \text{ then } X=x \text{ with probability } p \quad (4)$$

という簡単なルールを生成する．健康データマイニングでは、Y は生活習慣データ項目で y はその条件、X は健康データ項目で x はその値である．結論部は関心ある健康データ項目１つに制限するが、前提部を複数にすることを許容する．即ち、Z を別の生活習慣データ項目、z をその条件として、

$$\text{If } Y=y \text{ and } Z=z, \text{ then } X=x \text{ with probability } p \quad (5)$$

なるルールを許容する.ルール(5)をルール(4)の特殊化と呼ぶ．

ITRULE アルゴリズムでは、蓄積された多くのデータセットから有効なルールを生成するメカニズムとして式(6)に示す J 測度[10]を用いる．

$$J(x|y) = p(y)\left(p(x|y)\log\frac{p(x|y)}{p(x)} + (1-p(x|y))\log\frac{(1-p(x|y))}{(1-p(x))} \right) \quad (6)$$

式(6)において、大括弧内は $Y=y$ という事象が起きた場合に X の値に関して得られる情報の大きさを表す．ルールの場合には $X=x$ か $X=\bar{x}$ なので２項の和になっている．即ち大括弧内は $Y=y$ という前提がある場合とない場合で X の値に関する確率分布がいかに異なるかという尺度であるとも言える．J 測度はこれに母集団において $Y=y$ という事象が起きる確率 $p(y)$を掛けたもので、この値が大きいルールが良いルールということになる[10].

実際にはルールの J 測度は、蓄積されたユーザ毎の時系列データセットをサンプル集団として、$p(y)$をル

ールの前提条件がデータセットからのサンプルと一致する確率、$p(x)$をルールの結論がデータセットからのサンプルと一致する確率、$p(x|y)$を前提条件で条件付けられたルールの結論の条件確率、として計算する.

具体的なルールの自動生成プロセスは以下に述べるように、ユーザ毎の時系列データセットをサンプル集団とし J 測度が大きいルールが生き残るように実行される.

① 各出力フィールドX_i（関心のある健康データ項目）を順番に処理する．即ち、次の出力フィールドを対象とする前に、現在の出力フィールドに対してすべてのルールを生成する.

② 各出力フィールドX_iに対して、ある１つの値x_kを選択する．そして、次の値を対象とする前に、現在の値を結論とする全てのルールを生成する.

③ 各値x_kに対して、それぞれの入力フィールドY_jを選択する.

④ 各入力フィールドY_jに対して、それぞれの条件y_qを選択する．それぞれの条件は、入力フィールドのデータ型によって異なる.

a) シンボル値フィールドの場合、フィールドの各値が条件となる.

b) 数値型フィールドの場合、値はソートされ、各値が２分割境界としてテストされる．具体的には、各分割に対して J 測度が算出され、最も大きい J 測度を持つ分割が選択される．従って、「選択された分割値より大きい」と「選択された分割値以下」の２つの条件のみが可能となる.

⑤ ルール[$Y_j=y_q$ならば$X_i=x_k$]に対して、J測度を算出する.

⑥ 算出されたJ測度の値が、ルール格納テーブル中の同じ結論（$X_i=x_k$）を持つルールの中で最大のJ値より大きい場合、またはテーブル中のルール数が設定された最大数未満でかつ設定された最小サポート率および最小確信度基準を満たしていればテーブルにルールが格納され（必要に応じてJ値の低いルールが置き換えら)、さらに式(7)で示すJ_s値が評価される．それ以外の場合は、次の入力フィールドの条件に進む．ここで、サポート率とはサンプルデータセット中で、ルールの前提部が真のレコード数の割合、確信度とはルールの前提部が真のレコード中で、結論部が真のレコード数の割合である.

⑦ J_sの値が、ルール格納テーブル中のルールのJ値の最小値より大きい場合にはルール特殊化（前提条件の追加）を試みる.

⑧ すべての入力フィールド条件、入力フィールド、出力フィールド値、および出力フィールドを検討し終わるまで処理を繰り返す.

ここで、前述のJ_s値とルールの特殊化について説明する.ITRULEアルゴリズムではテーブルにルールが格納されると、そのルールを特殊化する（前提条件にさらに別の条件を追加する）潜在的な利点があるかどうか調べる．このために式(7)で定義されるJ_s値を評価する[10].

$$J_s = \max\left(p(y)p(x\mid y)\log\left(\frac{1}{p(x)}\right), p(y)(1-p(x\mid y)\log\left(\frac{1}{1-p(x)}\right)\right) \tag{7}$$

J_s値は特殊化を行った場合のJ値の上限値である．従って、この値がその時点でのルール格納テーブル中で同じ結論部を持つルールにおける最小のJ値より大きい場合にはルールの特殊化を試みる．即ち、前提条件にさらに別の入力フィールドを追加して、元の特殊化されていないルールの場合と同様な処理（①～⑧）を実行する.そして、特殊化されたルールのJ値がその時点でのルール格納テーブル中の最小のJ値を超えていれば、そのルールを置き換える.

特殊化は、追加する入力フィールドがなくなるまで処理を続けることが出来るが、健康データマイニングでは、前提条件が多いあまりにも複雑なルールは適当ではないので、２次のオーダーで止める（即ち前提条件は最大２つとする）ことにした.

3. 健康データマイニングの実例

3.1. ボランティアユーザの蓄積データ

あるボランティアユーザが蓄積した日常の生活習慣データ項目と健康データ項目およびそれぞれのデータ型を表１に示す.

表１．ボランティアユーザが蓄積したデータ項目とデータ型

生活習慣データ項目（データ型）	健康データ項目（データ型）
運動による消費エネルギー（数値）	体重（数値）
食事による摂取エネルギー（数値）	体脂肪率（数値）
アルコール摂取量（シンボル値）	最大血圧（数値）
睡眠時間（数値）	最小血圧（数値）
睡眠の深さ（シンボル値）	脈拍数（数値）
ストレス（シンボル値）	

運動による消費エネルギー（kcal）はユーザが身に付けている歩数計とフィットネスクラブの記録から、食事による摂取エネルギー（kcal）は朝、昼、夜のメニューからユーザの判断でデータ登録を行っている.

アルコール摂取量、睡眠の深さ、ストレスはユーザの判断でそれぞれ５段階（飲み過ぎ、多い、適度、少ない、非常に少ない）、３段階（ぐっすり、やや浅い、あまり眠れなかった）、３段階（多い、普通、少ない）のシンボル値でデータ登録している．

血圧（mmHg）と脈拍（回/分）は、オシロメトリック法による自動血圧計を用い、家庭でほぼ毎日朝同じ条件で測定している．血圧は測定のエラーを少なくするために、3 回計測しその平均をデータ登録している．

体重（kg）と体脂肪率（%）は、生体インピーダンス法による体脂肪率計を用い、やはり家庭でほぼ毎日朝同じ条件で測定しデータ登録している．

3.2. 入力変数（フィールド）の自動定義

ボランティアユーザの関心事は、日常の血圧と体脂肪に与える生活習慣の影響にあった．従って、出力変数は最大血圧（心臓収縮期血圧）と最小血圧（心臓拡張期血圧）および体脂肪率である．そこで、表１の生活習慣データ項目とこれらの健康データ項目の時系列データを基に 2.3.1 で述べた手法により入力変数（フィールド）を定義した．この時、ピアソンの積率相関係数はそれぞれのデータ項目について時系列の最初の 80 データを基に計算し、相関が有意であると判断できる相関係数の基準を 0.3 と仮定してR_s=0.3 と設定した．結果を表２にまとめた．

表２．計算された相関係数と各出力変数に対して定義された入力変数（フィールド）

A．対象出力変数：最小血圧

入力変数	r	定義された入力フィールド
消費エネルギー	-0.333	e_i+e_{i-1}+・・・+e_{i-5}(i=n-1)
摂取エネルギー	0.337	e_i(i=n-2)
アルコール摂取量	0.377	e_i(i=n-2)
ストレス	0.330	e_i(i=n-1)
実効睡眠時間	-0.312	e_i+e_{i-1}+・・・+e_{i-6}(i=n-1)

B．対象出力変数：最大血圧

入力変数	r	定義された入力フィールド
消費エネルギー	-0.309	e_i+e_{i-1}(i=n-1)
摂取エネルギー	0.345	e_i(i=n-2)
アルコール摂取量	0.322	e_i(i=n-2)
ストレス	0.310	e_i(i=n-1)

C．対象出力変数：体脂肪率

入力変数	r	定義された入力フィールド
消費エネルギー	-0.369	e_i+e_{i-1}+・・・+e_{i-4}(i=n-2)
摂取エネルギー	0.321	e_i+e_{i-1}+・・・+e_{i-4}(i=n-2)
アルコール摂取量	0.608	e_i(i=n-2)

表２において、rは 2.3.1 で述べたパラメータセット$((n\text{-}m)_{\max}, (i\text{-}j)_{\max}, s_{\max})$におけるピアソンの積率相関係数の値で、このパラメータセットに基づき入力フィールドは定義されている．

表２A における実効睡眠時間は睡眠時間を睡眠の深さの値（1（ぐっすり）、2（やや浅い）、3（あまり眠れなかった））の平方根で割って重み付けをしたものである．なお重み付けの方法はこれに限ったことではない．

3.3. 決定木によるルール生成

3.3.1. 血圧と生活習慣

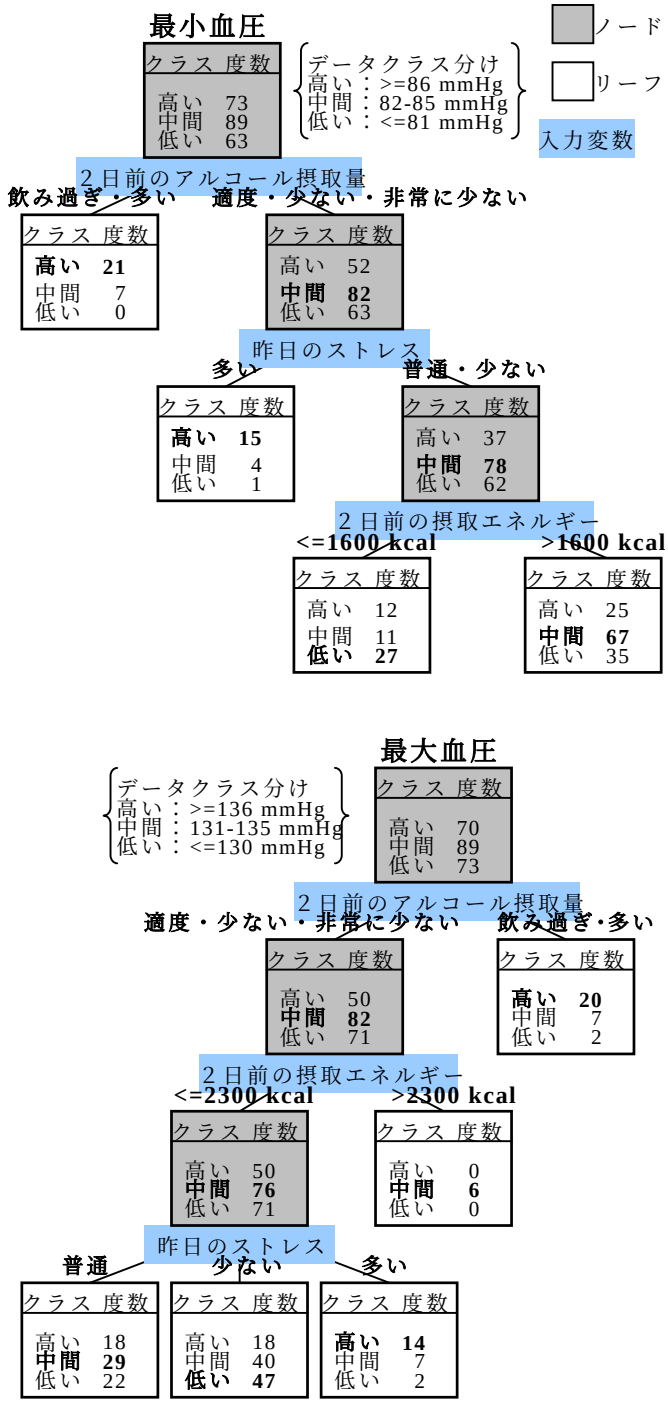

図４．最小血圧と最大血圧に関する決定木の例

C5.0 アルゴリズムで生成した最小血圧と最大血圧に関する決定木の例を図４に示す．サンプル集団のレコード数は約 230 である．このユーザの場合、２日前のアルコール摂取量と摂取エネルギーおよび昨日のストレスが、現在の血圧に影響を与えるキーの生活習慣データであることが判る．即ち、過度の飲酒が２日後の朝の血圧を上昇させ、前日の強いストレスが翌朝の血圧を上昇させ、少な目の食事摂取が２日後の朝の血圧を下げる傾向にあることが判る．

3.3.2. 体脂肪と生活習慣

C5.0 アルゴリズムで生成した体脂肪率に関する決定木の例を図５に示す．

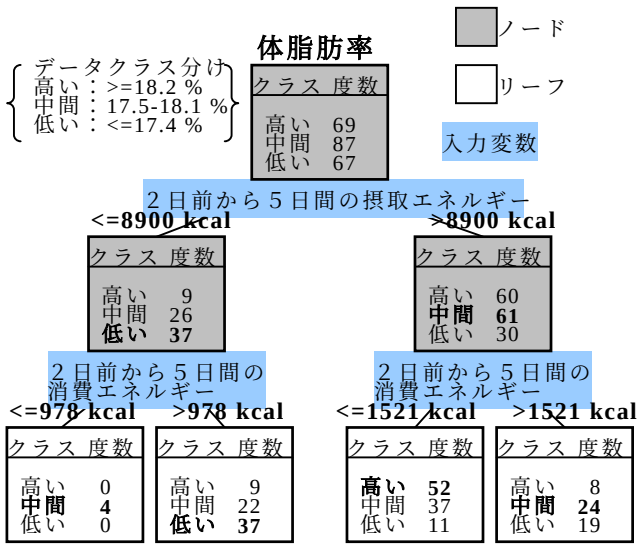

図５．体脂肪率に関する決定木の例

このユーザの場合、２日前から５日間の合計の摂取エネルギーと消費エネルギーが、現在の体脂肪率に影響を与えるキーの生活習慣データであることが判る．即ち、５日間の多目の食事摂取が２日後に体脂肪率の増加となって現れ、５日間の継続的なエネルギー消費が２日後に体脂肪率の減少となって現れる傾向があることが判る．血圧と異なり、５日間という継続的な生活習慣が現在の体脂肪率に影響を与えていると言える．

3.4. アソシエーションルール解析によるルール自動生成

3.4.1. ルール自動生成の流れ

ITRULE アルゴリズムを用いたアソシエーションルール解析の自動化を試みた．データマイニングサーバでの処理の流れを図６に示す．

１）データチェック

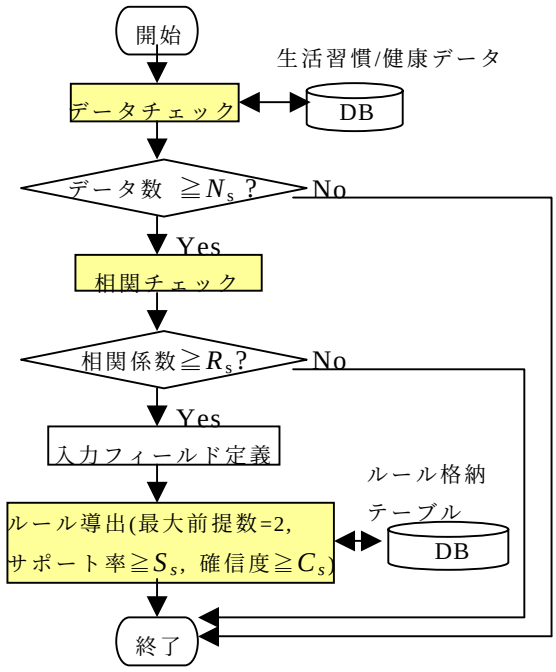

図６．ルール自動生成の流れ図

このプロセスは、各ユーザ毎に、どの程度の生活習慣データと健康データが蓄積されたかをチェックする．ユーザがデータ蓄積を開始して、最初の３ヶ月間でデータ数がある閾値N_sを超えていると、健康データマイニング対象ユーザとして登録し、自動的に入力フィールド定義、ルール生成プロセスを実行する．ここでは、10%程度のミッシングデータが許容されると仮定し、N_s=80 と設定した．

２）相関チェック

このプロセスは、*n-m*, *i-j*, *s*をパラメータとして変化させ（*n-m*=1~10, *i-j*=0~9, *s*=1~3）、最初の３ヶ月間蓄積された時系列データをもとに、各ユーザ毎にΔh_{nm}と$e^t{}_{ij}$の間のピアソンの積率相関係数*r*を計算する．もし、*r*の最大値がある閾値R_sを超えていたら、対応する$e^t{}_{ij}$が自動的に入力変数（フィールド）として定義される．ここでは、80~90のデータ数において相関が有意であると判断できる相関係数の基準を0.3と仮定し、R_s=0.3と設定した．

３）ルール生成

このプロセスは、2.3.3で述べた手法により生活習慣と健康状態の相関ルールを自動生成する．あまりにも複雑なルールはユーザを混乱させ、生活習慣の改善や健康増進を行うための指針にならないと考え、前提条件は２つまでに制限した．ルールの最小サポート率S_sと確信度C_sは可変で、ここでは、それぞれの出力変（健康データ項目）について生成されるルール数が10個以

下になるようにS_s=0.04、C_s=0.60~0.65に設定した.

3.4.2. 生成されたルール

決定木を生成した同じユーザの約１年間にわたり蓄積された時系列データ（データセット数 330~340）をもとに、最小血圧、最大血圧、体脂肪率についてルールを自動生成した.

１）血圧

S_s=0.04、C_s=0.65 と設定して自動生成された最小血圧に関する８個のルールを図７に、最大血圧に関する５個のルールを図８にまとめた. ここで、インスタンスIはルールの前提が真のレコード数、サポート率SはIを総レコード数で割った値、確信度Cはルール全体が真のレコード数をIで割った値である. ルールはS×Cの値が大きい順にソートして表示している. ここでルール４はルール３の特殊化で確信度がさらに高くなっている.

データセット数は異なるが決定木から得られる情報とほぼ同質のルールが得られており、２日前のアルコール摂取量と摂取エネルギーおよび昨日のストレスが朝の血圧に大きな影響を与えていることが判る. これらのルールは個人毎に異なると考えられ、ユーザはこのような個人特有の情報に基づいて個々の生活習慣の改善を行うことができる.

２）体脂肪率

S_s=0.04、C_s=0.60 と設定して自動生成された体脂肪率に関する４個のルールを図９にまとめた. やはり決定木から得られる情報と同質のルールが得られている. 特に最も確信度が高いルール２は有用で、このユーザの場合、５日間の平均消費エネルギーが 280 kcal以下の時に同じ５日間の平均摂取エネルギーが 2000 kcal を超えていると体脂肪率が増加する傾向があることが判る. 従って、ユーザは具体的な数値目標をもって体脂肪率の管理に取り組むことができる.

なお、２日前のアルコール摂取量が、相関係数の大きな入力フィールドとして選ばれている（表２C）が、ここでのルール生成条件においてはアルコール摂取量は前提条件に含まれないという結果になっている.

4. 考察

4.1. 健康データマイニングの自動化プロセス

健康データマイニングにおいては、それぞれの出力変数（健康データ項目）に対する入力変数（フィールド）の選定が最もキーとなるプロセスである. 全ての生活習慣データ項目の、全ての*i-j, s*の組み合わせを入力フィールドとしてデータマイニングを実行することは、個人健康管理システムのユーザが不特定多数であることを考慮すると、あまりにも膨大な計算量になる. またこのようにして得られるルールは複雑で判り難いものになると予想される.

出力フィールド: 最小血圧
レコード総数: 333
高い: ≧ 86 mmHg　　I：インスタンス
中間: 82-85 mmHg　　S:サポート率
低い: ≦ 81 mmHg　　C: 確信度
ルール導出条件: 最大前提数 = 2, S≧0.04, C≧0.65

	前提部	結論部	I	S	C
1	[ストレス1=少ない, 摂取1< 1625 kcal]	[最小血圧=低い]	40	0.12	0.7
2	[酒量1=飲み過ぎ]	[最小血圧=高い]	36	0.108	0.72
3	[ストレス1=多い]	[最小血圧=高い]	35	0.105	0.69
4	[ストレス1=多い, 消費6< 1819 kcal]	[最小血圧=高い]	27	0.081	0.81
5	[酒量1=飲み過ぎ, 摂取1>2025 kcal]	[最小血圧=高い]	22	0.066	0.86
6	[ストレス1=少ない, 酒量1=飲み過ぎ]	[最小血圧=高い]	23	0.069	0.78
7	[ストレス1=普通, 睡眠7> 41.67 時間]	[最小血圧=中間]	16	0.048	0.75
8	[酒量1=少ない, 摂取1> 1945 kcal]	[最小血圧=中間]	16	0.048	0.69

・ストレス1: 昨日のストレス
・摂取1: 2日前の摂取エネルギー
・酒量1: 2日前のアルコール摂取量
・消費6: 昨日から6日間の消費エネルギー
・睡眠7: 昨日から7日間の実効睡眠時間

図７. 最小血圧に関して自動生成されたルール

出力フィールド: 最大血圧
レコード総数: 340
高い: ≧ 137 mmHg　　I：インスタンス
中間: 132-136 mmHg　　S:サポート率
低い: ≦ 131 mmHg　　C: 確信度
ルール導出条件: 最大前提数 = 2, S≧0.04, C≧0.65

	前提部	結論部	I	S	C
1	[酒量1=飲み過ぎ]	[最大血圧=高い]	37	0.109	0.73
2	[ストレス1=少ない, 摂取1< 1625 kcal]	[最大血圧=低い]	40	0.118	0.65
3	[酒量1=飲み過ぎ, 摂取1< 2175 kcal]	[最大血圧=高い]	26	0.077	0.81
4	[ストレス1=少ない, 酒量1=飲み過ぎ]	[最大血圧=高い]	23	0.068	0.78
5	[ストレス1=多い, 消費2< 508 kcal]	[最大血圧=高い]	20	0.059	0.8

・ストレス1: 昨日のストレス
・摂取1: 2日前の摂取エネルギー
・酒量1: 2日前のアルコール摂取量
・消費2: 昨日から2日間の消費エネルギー

図８. 最大血圧に関して自動生成されたルール

出力フィールド: 体脂肪率
レコード総数: 331
高い: ≧ 18.0 %　　I：インスタンス
中間: 17.3-17.9 %　　S:サポート率
低い: ≦ 17.2 %　　C: 確信度
ルール導出条件: 最大前提数 = 2, S≧0.04, C≧0.60

	前提部	結論部	I	S	C
1	[摂取5> 9925 kcal]	[体脂肪率=高い]	37	0.112	0.62
2	[摂取5> 9925 kcal, 消費5< 1417 kcal]	[体脂肪率=高い]	29	0.088	0.72
3	[消費5> 1747 kcal]	[体脂肪率=低い]	21	0.063	0.62
4	[摂取5< 8225 kcal]	[体脂肪率=低い]	14	0.042	0.71

・摂取5: 2日前から5日間の摂取エネルギー
・消費5: 2日前から5日間の消費エネルギー

図９. 体脂肪率に関して自動生成されたルール

そこで、ルール生成プロセスの前処理として、３ヶ月間の時系列データをもとにした相関チェックにより、入力フィールドにある程度ふるいをかけた．即ち、各健康データ項目について、生成されるルールの前提部となる候補を予め“相関係数”という尺度を使って絞り込んだ．しかし、Δh_{nm}とより大きい相関係数をもつ$e^t{}_{ij}$が必ずしもより有効な入力フィールドとはかぎらない．何故ならば、ルールは因果関係であり相関関係とは異なるからである．すでに指摘したように２日前のアルコール摂取量は体脂肪率と大きな相関を示す（r=0.608）が、体脂肪率に関するルールの前提条件には入ってこない．この場合は問題ないが、相関係数が閾値R_sより小さいが有効な入力フィールドを見逃す可能性がある．相関が非線形の場合、みかけの相関係数が小さく計算されることがあるからである．未だここで述べたケースしかないが､今後の検討課題である．

4.2. 定義された入力フィールドの寿命

本研究のデータマイニング実行例では、入力フィールドを最初の３ヶ月間のデータセットをもとに自動定義し、ルール生成は３ヶ月毎に１回と設定した．ここで、日数とともにデータ数は増えていくが、定義した入力フィールドはどの位の期間有効であろうか．ユーザは、生成されたルールに基づいて生活習慣を改善するかもしれないし、同時に加齢を重ねる．従って、入力フィールドは時間によって変化するものと捉えるべきである．本研究のシステムでは、暫定的に入力フィールドは１年間有効とし、１年経つとまた最新の３ヶ月間のデータを基に入力フィールドを再定義するようシステムをデザインした．

4.3. システムの最適化

健康データマイニングの自動化システムでは、システム管理者は以下の４つのパラメータを調整することが出来るようにした．

N_s：３ヶ月間の蓄積データ数がこの値を超えていると、データマイニングプロセスを開始する．

R_s：入力フィールドが定義されるための最小のピアソン積率相関係数．

S_s：ルール格納テーブルに格納されるための最小のサポート率．

C_s：ルール格納テーブルに格納されるための最小の確信度．

今回の研究では、N_s=80、R_s=0.3、S_s=0.04、C_s=0.6~0.65 に設定して、あるボランティアユーザにとっていくつかの有用なルールを得ることができた．今後、できるだけ多くの不特定ユーザが満足できるルールが生成されるように、これらのパラメータを最適化する必要がある．

5. おわりに

筆者らが提案した個人健康管理のための健康データマイニングシステムが、システムのボランティアユーザの生活習慣と血圧と体脂肪率の関係において、健康管理に有用なルールを抽出できることを示した．健康データマイニングは、決定木およびアソシエーションルール解析に、本研究で開発した入力フィールドの自動定義アルゴリズムを組み合わせることで実現した．今後、多くのボランティアユーザを募り、システムを最適化することが課題である．

謝辞

本研究は、文部科学省科学研究費補助金制度基盤研究および文部科学省（科学技術振興事業団）地域研究開発促進拠点支援事業（RSP 事業）の支援を受けている．本研究は、また日本データベース学会および日立製作所による HiRDB アカデミックプログラムの支援を受けた．

文　献

[1] N. H. Lovell, F. Magrabi, B. G. Celler, K. Huynh, and H. Garsden, “Web-based acquisition, storage, and retrieval of biomedical signals,” IEEE Eng. Medicine and Biology, vol.20. no.3, pp.38-44, 2001.

[2] J. Cai, S. Johnson, and G. Hripcsak, “Generic data modeling for home telemonitoring of chronically ill patients,” Proc. AMIA Symp. pp.116-120, 2000.

[3] C. Mazzi, P. Ganguly, and M. Kidd, “Healthcare application based on software agents,” Medinfo 2001 Proceedings, pp.136-140, 2001.

[4] T. Hashiguchi, H. Takeuchi, and A. Uemura, “Highly advanced healthcare support services for the 21st century,” Hitachi Review, vol.50, no.1, pp.2-7. 2001.

[5] 竹内裕之, 橋口猛志, 新谷隆彦, “日常の健康管理を目的とした個人対応動的データベース” 医療情報学 vol.23, no.6, pp.497-502, 2004.

[6] H. Takeuchi, T. Hashiguchi, and T. Shintani, “Personal dynamic healthcare system utilizing mobile phone and web technologies,” Proc. 2nd Int. Conf. on Advances in Biomedical Signal and Information Processing, pp.304-307, 2004.

[7] H. Takeuchi, N. Kodama, T. Hashiguchi, and N Mitsui,” Healthcare data mining based on a personal dynamic healthcare system,” Proc. 2nd Int. Conf. on Computational Intelligence in Medicine and Healthcare, pp.37-43, 2005.

[8] M. F. Usama, P. S. Gregory, P. Smyth, and Ramasamy, Advances in Knowledge Discovery and Data Mining, The AAAI Press, 1996.

[9] M. J. A. Berry and G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support, John Wiley & Sons, Inc., 1997.

[10] P. Smyth and R. M. Goodman, “An information theoretic approach to rule induction from databases,” IEEE Trans. Knowledge and Data Engineering, vol.4, no.4, pp.301-316, 1992.