健康データマイニングにより自動抽出されたルールの妥当性検証

竹内　裕之　　児玉　直樹

高崎健康福祉大学健康福祉学部医療情報学科 〒370-0033 群馬県高崎市中大類町 37-1

E-mail: {htakeuchi, kodama}@takasaki-u.ac.jp

あらまし　個人の日常の健康状態と生活習慣に関わる時系列データをサーバに蓄積し，両者の相関ルールを抽出する「健康データマイニング」をコア技術とした，クラウド型個人健康管理システムを開発している．「健康データマイニング」では，クラウドで自動的にルールを抽出するために，前処理として健康状態と生活習慣の遅延期間を考慮した時系列データ間の相関分析を行っている．本稿では，ボランティアユーザについて自動抽出された相関ルールを散布図に立ち戻って検証し，開発中の「健康データマイニング」アルゴリズムの今後の課題について考察した．

キーワード　個人健康管理システム 健康データマイニング 時系列データ解析

1. はじめに

インターネットを活用した健康医療分野のユビキタス化が進展しており，最近の国際学会の潮流としても，m(mobile)-health や p(personalized)-health といった概念が浸透している[1,2]．特に体重，体脂肪率，血圧といった個人の健康に関するデータが家庭や職場などでも容易に取得できるようになり，携帯電話やスマートフォン等の携帯端末を通してインターネット上（クラウド）に蓄積できる技術が開発されている．我々はいち早く，クラウドで処理を行う自動健康データマイニングをコア技術とした個人健康管理システムを開発してきた[3,4]．このシステムは，携帯端末を通して入力した個人の日常の生活習慣と健康に関するデータをクラウドに蓄積し，生活習慣と健康状態の相関ルール抽出（健康データマイニング）を行い，その結果を個人の携帯端末から参照できるものである．クラウド型であるため，保健師や管理栄養士など保健指導者が参画する運用も可能である[5]．本システムでは，個人の生活習慣や健康に関するデータを日毎の粒度で時系列的に蓄積することを前提としており，健康管理を行う多くの人々の長期にわたるデータはまさに健康ビッグデータを構成する．今後，ウェアラブルセンサーなどから発生するストリーム状の生体情報を扱うようになるとそのデータ量はさらに膨大なものとなる．

本稿では，本学の学生を中心とした個人健康管理システムのボランティアユーザが，2012 年 6 月 1 日から 11 月 30 日までの 6 か月間に日毎の粒度で蓄積した生活習慣と健康に関するデータに基づき，開発した健康データマイニング手法によって得られた，パターンやルールについてその妥当性を検証する．

2. 研究方法

2.1. 対象ユーザ

本研究の対象ユーザはいずれも本学の学生であり，22 歳の男女 3 名である．ユーザ A(女)は，特定保健用食品である健康茶の摂取が，ユーザ B(女)は，豆乳摂取が，それぞれのユーザの日々の体重・体脂肪率に与える影響を調べた．またユーザ C(男)は喫煙の習慣があり，喫煙量と血圧の関係を調べた．

2.2. データの取得方法

体重，体脂肪率は，タニタの体組成計（Inner Scan: BC-521）を用い，毎朝起床後もしくは毎日入浴前（ユーザによって異なる）に計測した．血圧，脈拍数はオムロン社の自動血圧計（オシロメトリック法）を用いて起床後に計り，血圧については 3 回計測してその平均値をデータ登録した．

生活習慣としての消費エネルギーは，歩行によるものはオムロンの歩数計（Walking style）を携帯して計測し，その他の運動については Mets 値を基に推測した．必要に応じて，摂取エネルギーについては毎食事の内容からインターネット上の関連サイトを参照するなどして推測した．また，健康茶，豆乳などの摂取量は 1 日あたりの摂取量（ml）を記録した．喫煙量は 1 日当たりの喫煙量（タバコ本数）を記録した．

2.3. 健康データマイニングの概要

我々が開発している健康データマイニングでは，「生活習慣の蓄積が健康状態に変化をもたらし，その影響は時間遅れをもって現れることがある」という極めてシンプルなモデルをベースとしている[6]．すなわち，ある健康状態の変化を出力変数とし，時間遅れを考慮したある期間の生活習慣の蓄積を入力変数として相関ルールを抽出する．相関ルールの抽出には，まず入力変数のいたずらな増加を防ぐために，あらかじめ

時系列データを基にして，出力変数である健康状態の変化に影響を及ぼす生活習慣の蓄積をスクリーニングする．スクリーニングには式(1)で表される時系列データ間のピアソンの積率相関係数を用いる．

$$r(\Delta h_{nm}, e^t{}_{ij}) = \frac{Cov(\Delta h_{nm}, e^t{}_{ij})}{SD(\Delta h_{nm})SD(e^t{}_{ij})} \qquad (1)$$

ここで，

$$\Delta h_{nm} = h_n - h_m \qquad (2)$$

は目的変数である健康データ h の変化を表す差分値であり，

$$e^t{}_{ij} = e_i + e_{i-1} + \cdots + e_j \qquad (3)$$

は生活習慣データ e の蓄積を表す，ある期間に亘る加算値である．時間遅れは遅延期間 $s = n - i \geqq 1$ で表現する（図 1 参照）．

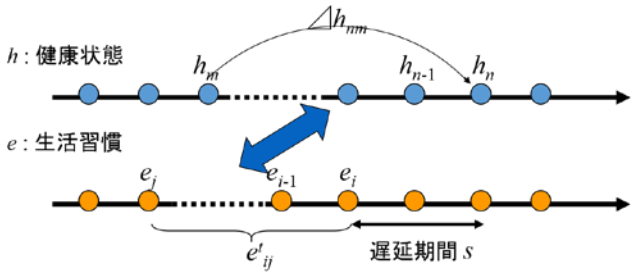

図 1　健康状態と生活習慣の時系列相関の評価

式(1)において，r は相関係数，$SD(\Delta h_{nm})$はΔh_{nm} の時系列区間（ここでは 3 ヶ月）における標準偏差，$SD(e^t{}_{ij})$ は $e^t{}_{ij}$ の時系列区間（ここでは 3 ヶ月）における標準偏差，$Cov(\Delta h_{nm}, e^t{}_{ij})$は$\Delta h_{nm}$ と $e^t{}_{ij}$ の共分散である．

実際のスクリーニングは，対象とする健康状態 h と各種生活習慣 e の時系列データについて，n-m, i-j, s をパラメータとして式(1)のピアソンの積率相関係数を評価する．すなわち，各(n-m, i-j, s)のセットにつき式(1)を評価し，もし，1 つ以上の相関係数の絶対値がある閾値 r_s より大きければ，その e の蓄積を h に対する入力変数として採用する．そして，ルールマイニング処理における実際の入力フィールドは相関係数の絶対値が最大となる(n-m, i-j, s)のセット$((n\text{-}m)_{\max}, (i\text{-}j)_{\max}, s_{\max})$をもとに定義する．例えば，$(i\text{-}j)_{\max}$=2, $s_{\max}$=2 であれば，e に関わる入力フィールドを

$$e_i + e_{i-1} + e_{i-2} \quad (i=n\text{-}2) \qquad (4)$$

と定義する．すなわち，「2 日前から 3 日間の生活習慣 e の蓄積」を入力変数のひとつとして定義する．ここで，$(i\text{-}j)_{\max}$ が大きいということは，長期間の生活習慣の蓄積が現在の健康状態に影響を与え，$s_{\max}$ が大きいということは，生活習慣の蓄積が遅れをもって現在の健康状態に影響を与えるということになる[7]．

次に，時系列相関によるスクリーニングで採用された生活習慣の蓄積を入力変数 Y（通常複数）とし，対象とする健康状態を，その時系列データが「高い」「中間」「低い」の 3 つのシンボル値を持つ出力変数 X として，ITRULE アルゴリズム[8]を用いた相関ルールマイニングを行う．ITRULE アルゴリズムは，

If Y=y, then X=x with probability p

という相関ルールを生成する．このアルゴリズムでは，多くのデータセットから有効な相関ルールを抽出するために，式(5)で表される J 測度を用いてルールを評価する[8]．

$$J(x|y) = p(y)\left(p(x|y)\log\frac{p(x|y)}{p(x)} + (1-p(x|y))\log\frac{(1-p(x|y))}{(1-p(x))} \right) \qquad (5)$$

J 測度は $Y=y$ 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量の大きさ，つまり $Y=y$ という前提がある場合とない場合で X の値 x に関する確率分布がいかに異なるかという尺度に，$Y=y$ という事象が起きる確率 $p(y)$を掛けたものであり，この値が大きいほどよい相関ルールということになる．

3. 自動抽出されたルールとその検証

3.1. 健康茶摂取に関わるルール

ユーザ A は 22 歳女性で，特定保健用食品となっている健康茶摂取の効果を検証しようとした．このユーザは摂取のタイミングに関心を持ち，当初 3 カ月間は，食前，食中，食後と摂取のタイミングを変えてデータを取得した結果，相関解析で食中の摂取が有効であることを見出した．そこで，その後の 3 カ月間は食中摂取にタイミングを絞り，摂取量を変化させながら毎日健康茶を摂取した．体重，体脂肪率は毎日入浴前に計測した．

サーバで自動的に実行された健康データマイニングからは，

「4 日間の総摂取カロリーが 1 日平均 1369.5 kcal より大きい，かつ 2 日間の健康茶摂取量が 1 日平均 177.5

ml 未満ならば 2 日後の体脂肪率が高い傾向にある」

[確信度：77.8%　サポート率：10.5%]

というルールが抽出された．このユーザとしては，食事による摂取カロリーが 1 日平均 1370 kcal より多く，健康茶の摂取量が 1 日平均 178 ml より少ないと，体脂肪率が高くなるということで，生活習慣上の目標が得られたことになる．

ルールの妥当性を検証するために，まず 4 日間の 1 日平均総摂取カロリーと 7 日前からの体脂肪率変化の散布図を図 2 に示す．データ数 n = 66，相関係数 r = 0.488，1%水準で正の相関が見られ，1 日平均摂取カロリーが 1,350 kcal を超えると，すべての体脂肪率変化がプラスになっている．次に，2 日間の 1 日平均健康茶摂取量と 7 日前からの体脂肪率変化の散布図を図 3 に示す．データ数 n = 67，相関係数 r = -0.356，1 %水準で有意な負の相関がみられ，1 日平均健康茶摂取量が 180 ml 未満であると体脂肪率変化がプラスになる確率が高いことが判る．このように，ここで自動抽出されたルールは散布図から裏付けられている．

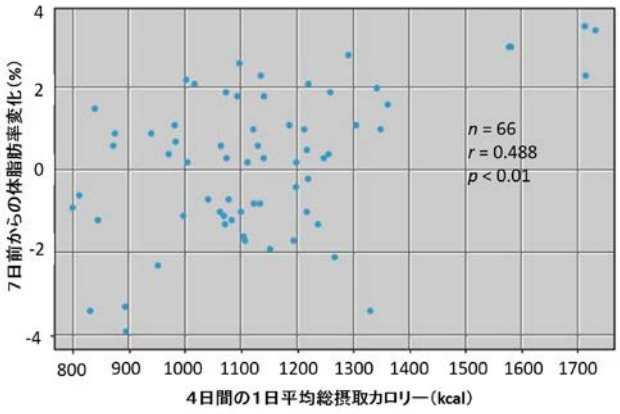

図 2　1 日平均総摂取カロリーと 7 日前からの体脂肪率変化の散布図

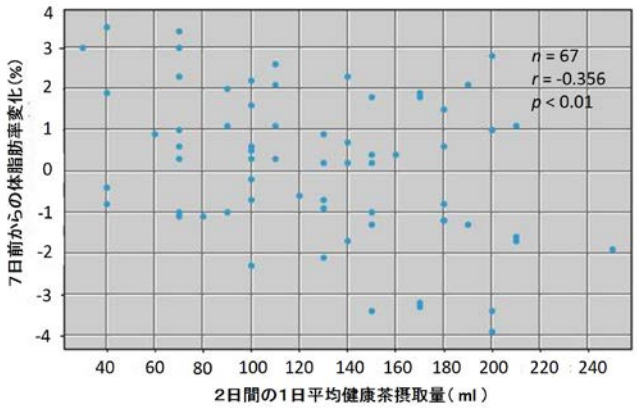

図 3　1 日平均健康茶摂取量と 7 日前からの体脂肪率変化の散布図

3.2. 豆乳摂取に関わるルール

ユーザ B は 22 歳女性で，豆乳に含まれる大豆たんぱく質などのダイエット効果を検証しようとした．大豆たんぱく質に含まれるβ―コングリシニンは，「摂取した油脂を完全には消化せず，一部を未消化で体外に排出する」作用があるとされ，体重，体脂肪率の低下が期待された．しかし当初 3 カ月間のデータからは，

「10 日間の豆乳摂取量が 1 日平均 550 ml より多いならば，3 日後の体重が高い傾向にある」

［確信度：100%　サポート率：15.0%］

というルールが自動抽出された．該当する散布図（図 4)を検証すると，体重変化は豆乳摂取量と正の相関を示し，確かに 1 日平均摂取量が 540 ml を超えるあたりから殆どの体重変化がプラスであることが判る．

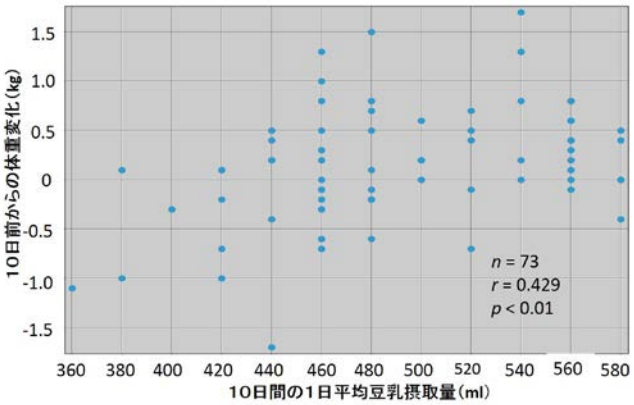

図 4　1 日平均豆乳摂取量と 10 日前からの体重変化の散布図

しかしこのユーザは，豆乳摂取量が 420 ml 以下であると体重変化は殆どマイナスであることに気がついた．さらに図 5 に示したように体脂肪率変化と豆乳摂取量の散布図が非線形であり，豆乳摂取量が 480 ml 以下であれば体脂肪率も低下することが判った．

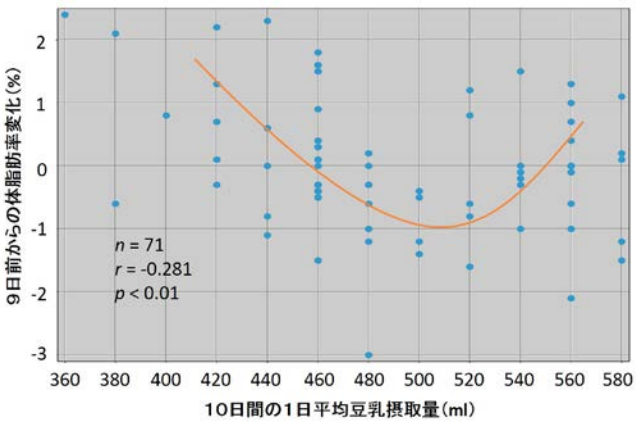

図 5　1 日平均豆乳摂取量と 9 日前からの体脂肪率変化の散布図

そこで，このユーザは，その後の３カ月間は１日当たりの豆乳摂取量を最多 400 ml に制限してデータを蓄積し，体重，体脂肪率変化との相関をみた．体重，体脂肪率は起床後朝食前に毎日計測した．その結果，

「10 日間の豆乳摂取量が１日平均 305 ml より多いならば，翌日の体重が低い傾向にある」

［確信度：86%　サポート率：8.6%］

というルールが自動抽出された．

図６は，ルールを検証するための，後半３カ月間の体重と，摂取量を制限した10日間の１日平均豆乳摂取量の散布図である．データ数 n = 72，相関係数 r = −0.700，１％水準で確かに負の相関が得られた．さらに，　体脂肪率と 10 日間の１日平均豆乳摂取量の間にも，5%水準ではあるが負の相関がみられた（図 7）．

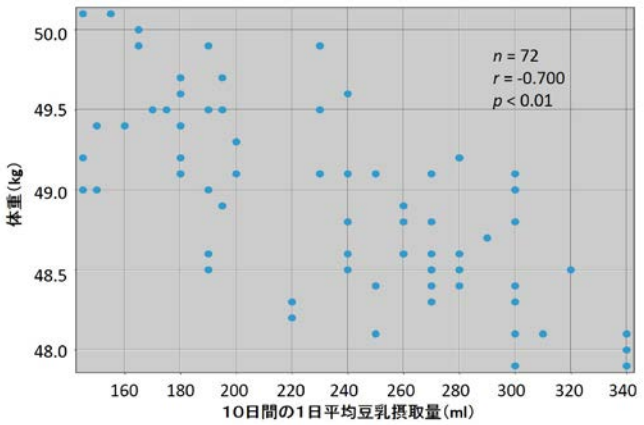

図６　１日平均豆乳摂取量と体重の散布図

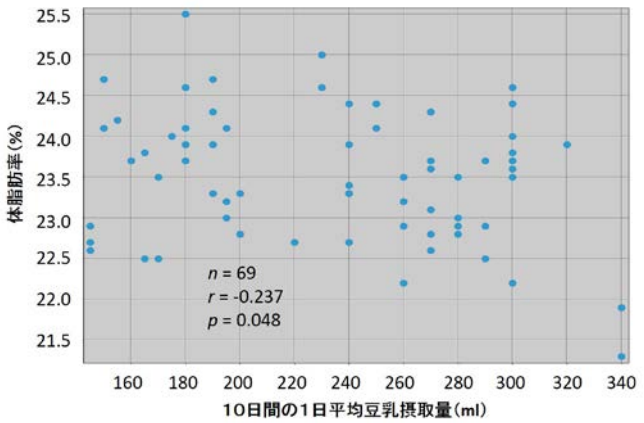

図７　１日平均豆乳摂取量と体脂肪率の散布図

3.3. 喫煙に関わるルール

ユーザ C は喫煙の習慣を持つ 22 歳の男性である．健康状態としては体重と血圧に関心があり，喫煙量との関係を検証した．血圧は起床後 30 分以内に毎日計測し，最大（心臓収縮期）血圧と最小（心臓拡張期）血圧を３回計測し，その平均値を記録した．計測には，オムロン社の自動血圧計（オシロメトリック法）を用いた．

自動実行された健康データマイニングからは，

「喫煙量が１日 8.5 本より多いならば，翌日の最大，最小血圧が高い傾向にある」

［確信度 = 83.3%　サポート率 = 20%］

というルールが抽出された．図８と９はこのルールを検証するための，１日の喫煙量と 2 日前からの最大，最小血圧変化の散布図である．確かに，１日の喫煙量が８本を超えると最大，最小血圧とも変化は殆どプラスになっていることが判り，散布図からルールの妥当性を検証できた．因みに，最大血圧ではデータ数 n = 82，相関係数 r = 0.529，1%水準で有意な正の相関が得られ，最小血圧でも 同じように 1%水準で有意な正の相関が得られている．

喫煙が血圧，脈拍数を急上昇させることはよく知られている[9]が，生活習慣としての喫煙が，起床後の安静時血圧にも影響していることが示唆された．

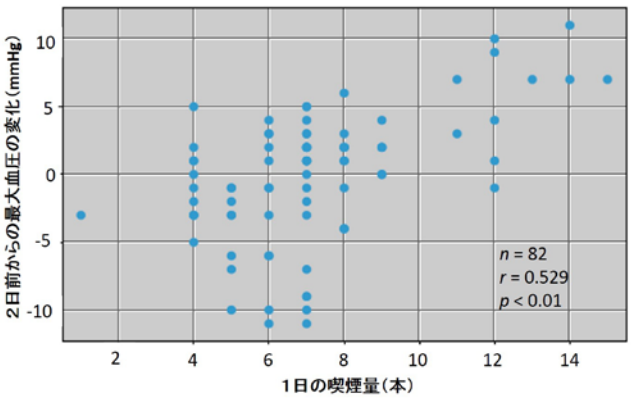

図８　１日の喫煙量と 2 日前からの最大血圧変化の散布図

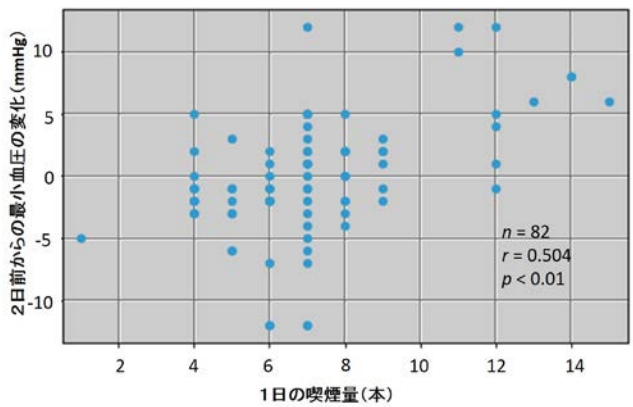

図９　１日の喫煙量と 2 日前からの最小血圧変化の散布図

4. 考察

4.1. 健康データマイニングのアルゴリズム

生活習慣データの蓄積と健康状態の変化の時系列相関係数の大きさ（ここでは 0.3 以上）でスクリーニングをかけて，ターゲットである健康状態に対して生活習慣データからなる入力変数を自動設定するのが，健康データマイニングのコアの部分である[7]. 相関ルールは因果関係であり，相関関係とは異なるのであるが，因果関係があるなら少なからず相関も認められるであろうという前提でこのアルゴリズムは成り立っている.

このような観点で 3 人のユーザについての結果を考察する．このうち最も有効に相関ルールが自動抽出されたのは，**3.3** の例である．喫煙量と最大，最小血圧の時系列相関係数は 0.3 を超えているので，喫煙量は健康データマイニングの入力変数として自動設定される．そして，図 8, 9 の散布図から判るように，1 日の喫煙量が 8 本以下であれば，血圧の変化はプラスとマイナスの確率はほぼ同等であるが 8 本を超えると 80%以上の確率でプラスになっている．健康データマイニングのアルゴリズムは迷うことなく因果関係であるルールを出力する.

3.1 の例も比較的わかりやすい．体脂肪率変化との関わりにおいて，摂取カロリーおよび健康茶摂取量との時系列相関係数がそれぞれ 0.3 を超えており，双方が入力変数として選ばれた．そして，図 2 の散布図をみると 1 日の平均総摂取カロリーが 1300 kcal 以下であると体脂肪率変化はプラスとマイナスの確率がほぼ同程度であるが，1300 kcal を超えると 80%以上の確率でプラス変化となる．図 3 の健康茶摂取量についての散布図では，目視で明瞭な因果関係はみられないが，アルゴリズム上特殊化の条件として相関ルールに反映されたと考えられる.

これらは，健康データマイニングが成功した例であるが，**3.2** の例は非線形性の強いケースで誤りではないが自動健康データマイニングがユーザに適切な情報を与えない場合である．事実は，10 日間にわたる豆乳過剰摂取はむしろ体重および体脂肪率を増加させるということであるが，自動抽出されたルールだけみると「豆乳摂取は体重を増加させる」ということになる．体脂肪率に関しては図 5 から判るように強い非線形相関が現れており，豆乳摂取量との間で単純に相関係数を求めるとその大きさは 0.3 より小さく入力変数に選ばれない．したがってルールも抽出されないという結果になっていた．結論としてダイエットを目的とした豆乳摂取量には適量があり，その範囲で摂取することにより期待した相関ルールが抽出され，散布図（図 6,7）からもそれが裏付けられた.

4.2. ルール自動抽出の今後の課題

生体における生活習慣と健康状態の関係は程度の差こそあれ本来非線形であり，線形を前提とした積率相関係数のみで因果関係へのスクリーニングをかけるにはリスクがある．特に，**3.2** の例のように非線形性が強い場合には，生活習慣データの範囲を限定してルールマイニングを実行する必要がある．今後相関の非線形性を自動認識して，非線形性が強い場合には生活習慣データの範囲を分割して，それぞれの分割範囲内でルール生成を行うなどの工夫が必要になる.

5. まとめ

本学の学生を中心とした個人健康管理システムのボランティアユーザが，2012 年 6 月 1 日から 11 月 30 日までの 6 か月間に日毎の粒度で蓄積した生活習慣と健康に関するデータに基づき，開発した健康データマイニング手法によって得られた，パターンやルールについてその妥当性を検証した.

その結果

（１） 本稿で対象とした３ボランティアユーザの内２名についてはサーバに蓄積された時系列データに基づき，開発した健康データマイニングのアルゴリズムによって，適切に生活習慣と健康状態の間の相関ルールが自動抽出されていた.

（２） 他の 1 名については，誤りではないが妥当ではない相関ルールが自動抽出され，その原因は生活習慣データと健康状態の間に強い非線形の相関があることによることが判った.

（３） 上記の場合，生活習慣データの範囲に制限を設けることにより適切に健康状態との間の相関ルールが抽出されることを確認できた.

謝辞

本研究は文部科学省科研費（課題番号：23500813）の助成を受けている．また，日本データベース学会と日立製作所による日立 HiRDB アカデミック制度の適用を受けている.

本研究に協力していただいた高崎健康福祉大学健康福祉学部医療情報学科の学生諸氏に感謝します.

参 考 文 献

[1] H. Kumpusch, D. Hayn, K. Kreiner, M. Falgenhauer, J. Mor, and G. Schreier, “A Mobile Phone Based Telemonitoring Concept for the Simultaneous Acquisition of Biosignals and Physiological Parameters”, Proc. 13th World Congress on Medical and Health Informatics

(Medinfo2010), pp. 1344-1348, 2010.
[2] E. C. Kyriacou, C. S. Pattichis, and M. S. Pattichis, ” An Overview of Recent Health Care Support System for eEmergency and mHealth Applications”, Proc. 31st Annual International Conference of the IEEE EMBS, pp. 1246-1249, 2009.
[3] H. Takeuchi, T. Hashiguchi, and T. Shintani, “Personal Dynamic Healthcare System Utilizing Mobile Phone and Web Technologies”, Proc. 2nd Int’l Conf. Advances in Biomedical Signal and Information Processing, pp. 304-307, 2004.
[4] 竹内裕之, 児玉直樹, 橋口猛志, 林　同文, “インターネット上で動く自動健康データマイニングシステム” 高崎健康福祉大学紀要　第 5 号 pp. 1-11, 2006.
[5] H. Takeuchi, Y. Mayuzumi, N. Kodama, and K. Sato, “Application Service Provider System for Healthcare with Data Mining Function”, Proc. 13th World Congress on Medical and Health Informatics (Medinfo 2010), 2010.
[6] 竹内裕之, 児玉直樹, “生活習慣と健康状態に関する時系列データ解析手法の開発”, DEWS 2008.
[7] 竹内裕之, 児玉直樹, 橋口猛志, 林　同文, “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”, DEWS 2006.
[8] P. Smyth and R. M. Goodman, “An Information Theoretical Approach to Rule Induction from Databases”, IEEE Trans. Knowledge and Data Engineering, vol.4, no.4 pp. 301-316, 1992.
[9] 上園慶子, 佐々木　悠, 川崎晃一, 浦江明憲, 天本敏昭, “喫煙の血圧・脈拍に対する急性効果とその喫煙時刻による差異”, 健康科学（九州大学） 第 15 巻 pp. 85-89, 1993.