個別化健康管理システムにおけるデータマイニング手法の研究

中島 緋沙恵† 竹内 裕之†

†高崎健康福祉大学大学院 医療福祉情報学専攻 〒370-0033 群馬県高崎市中大類町 37-1

E-mail: †{1610101, htakeuchi}@takasaki-u.ac.jp

あらまし 日本では，少子高齢化に加え医療費の高騰が進んでおり，保険財政が破綻する状況にある．医療費削減の為には個人レベルでの健康管理の普及が求められる．個人健康管理の普及には健康管理を支援するシステムが必要である．現在，健康管理を支援するシステムとして健康データマイニングの手法を用いた健康管理システムがある．しかし，健康維持のルールを作成するための説明変数を選択する際に，非線形性の高い人間のデータに対し相関係数を用いている問題点があった．よって，本研究では相関係数によらず説明変数を選択する手法を提案し，より良いルール抽出を行う手法の開発を目的とする．過去のデータマイニングの結果から良いルールが作成されると分かっている散布図の特徴に着目し，散布図に特徴をもつ説明変数を自動選択するための手法を提案する．提案手法により作成されたルールの評価は，リフト値によって従来手法により作成されたルールと比較し，提案手法の有用性を検討した結果、学生 12 名中 9 名において提案手法のリフト値が従来手法のリフト値よりも高い結果が得られた．よって，より的確な健康管理ルールを示すことができることから，有用であると考えられる．

キーワード 健康データマイニング,時系列データマイニング,遅延相関分析,個別化健康管理

1. はじめに

現在，わが国では少子高齢化が急速に進展しており，2015 年の時点で高齢化率が 26.7%となっている．今後も年少人口および生産年齢人口の割合は減少を続けると見込まれているため，高齢化率は今後上昇の一途を辿ると考えられる[1]. 厚生労働省の推定によると, 2060 年時点では約 2.5 人に 1 人が 65 歳以上の高齢者となる見込みであるとされている[2]．少子高齢化が進む中，医療費の肥大化も年々深刻になっており，大きな社会問題として広く認知されている．国民医療費に占める後期高齢者医療費の割合は平成 26 年度時点で 35.6%を占め，現在もなお増加傾向にある[1]．今後さらに少子高齢化が進む中で，医療費の高騰により保険財政が破綻する恐れがあると考えられる．

この問題を解決するために，個人ごとの健康管理が重要になってくると考えられる．しかし，自分の健康を管理するために何をしたら良いかを十分に理解している人は少ない．そのため，個人の健康管理を支援するシステムの普及が今後一層求められると考えられる．

現在，健康管理を支援するシステムとして竹内らによる健康データマイニングの手法を用いた個人健康管理システムが存在する[3-5].

健康データマイニングでは，遅延相関分析法によって相関係数の高い生活習慣データの蓄積を説明変数として選択しているが，相関係数が最大のものが必ずしも説明変数として適しているとは言えない. なぜなら，相関係数は線形的な関係を示す値であるが，健康状態などの人間から得られるデータが，線形的な変化を示すことは考えられないからである．よって，本来であればルール抽出に最適な説明変数として選択されるべき生活習慣の蓄積が，相関係数が小さいために説明変数として選択されていないと考えられる．

個人レベルでの健康管理の普及を進めるには，個人健康管理支援システムにおいて「何が健康に影響するのか」を簡潔かつより精度の高いルールとして示すことが重要である．よって，本研究では，従来手法の問題点を解決し，より良いルール抽出を行うデータマイニング手法の開発を目的とする．

本稿の構成は次の通りである．1 章では，我が国の現状と現行の個人健康管理システムの問題点，本研究の目的について述べた．2 章では健康データマイニングについて述べ，3 章では提案手法について述べる．4 章では実験および評価を行い 6 章で結論を述べる．

2. 健康データマイニング

個人健康管理システムに実装されているデータ解析技術であ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，その影響は時間の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[3]．健康データマイニングによるデータ解析では，個人が日々記録した生活習慣と健康状態の時系列データから個人ごとの相関ルールを生成する．健康データマイニングのコンセプトの概念図を図 1 に示す．

まず個人の現在の健康状態は，日常生活の影響を何らかの形で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，個人差も受けやすいものと想定する. 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ごとに時系列に蓄積し，その中から生活習慣と健康状態の間に何

らかの規則性を見出し個人ごとのルールとして抽出することである[4].

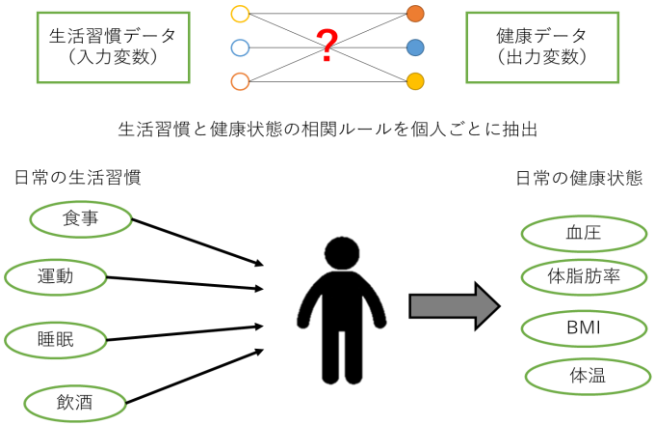

図 1　健康データマイニングのコンセプトの概念図

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（説明変数）Y, 健康データ項目を出力変数（目的変数）Xとして位置づけ,「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」といった相関ルールを個人ごとに抽出する．すなわち，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．健康データマイニングを実装したシステムのユーザはこのようなルールを健康管理や健康増進のために役立てることができる[5].

健康データマイニングに用いられている遅延相関分析法を図 2 に、健康データマイニングの処理フローを図 3 に示す.

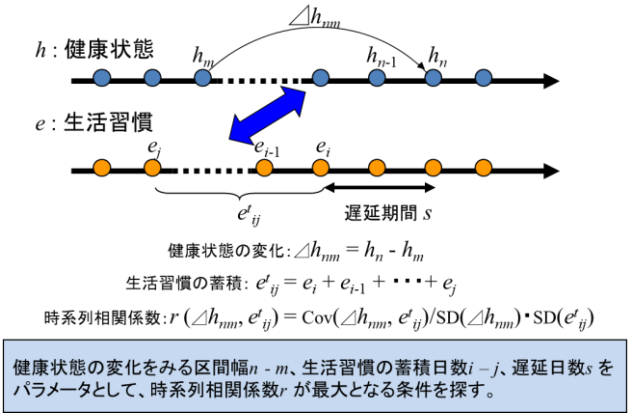

図 2　遅延相関分析法

健康データマイニングでは，次のように処理を行う.

1）　データチェック

このプロセスはユーザごとに 3 か月間の蓄積されたデータ数を登録されている項目ごとにチェックする．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

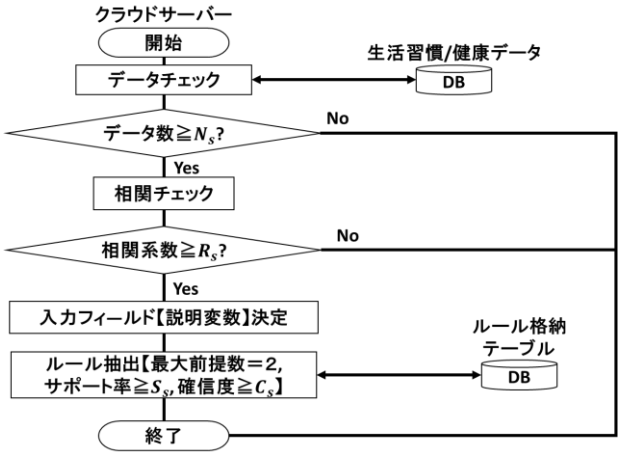

図 3　健康データマイニングの処理フロー

入力フィールド定義，ルール生成プロセスを実行する[4].

2）　入力フィールド定義

このプロセスは，健康データの変化（Δh_{nm}），生活習慣データの加算日数（e_{ij}^t），遅延日数をパラメータとして変化させ，Δh_{nm}とe_{ij}^tの相関係数 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，そのユーザに対してe_{ij}^tをΔh_{nm}に対する入力変数に定義する[4].

3）　ルール生成

このプロセスでは，ITRULE のアルゴリズムにより生活習慣と健康状態の相関ルールを生成する．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実施する指針にならないと考えられ，前提条件は 2 つまでに制限されている[4].

3. 提案手法

3.1. 提案手法のコンセプト

過去のデータマイニングを行った結果から，確信度が高いルール抽出ができたデータセットにおいて，健康データのプラス変化とマイナス変化に着目した時に，生活習慣データの蓄積がある一定の値を境に特異的な偏りを持つような散布図となることが分かっている．そのような散布図の簡易モデルを図 4 に示す.

この散布図において，ばらつきの多い範囲では特定の生活習慣だけでなく他の要因によって健康データが変化する一方で，偏りのある範囲では特定の生活習慣が健康データの変化に大きな影響を与えていると推察した．つまり，2 つの範囲の境界線が決定木分析によるルール抽出の際に，データを分類する明確な区切りとなっていると考えられる．ここで，散布図上でデータのばらつきが多い範囲と特異的な偏りのある範囲の

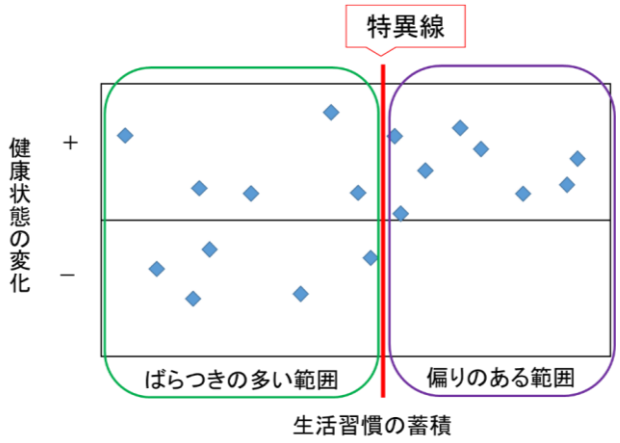

図 4　特異的な偏りを持つような散布図の簡易モデル

境界線を特異線と定義する．

特異線を持つような散布図の多くは，線形的な相関関係に乏しいため，相関係数によらず特異線に基づき説明変数を選択する手法では，ルール抽出の説明変数として適しているものの，従来手法では相関係数が低いために選択されなかった生活習慣データの蓄積も説明変数として選択することができると考えられる．よって，提案手法により選択された説明変数を用いることにより，従来手法よりも優れたルールの抽出ができると考えられる．

3.2. 特異線

特異線について，10 日前からの最高血流速度変化と 1 日の歩数の散布図を例に具体的に示す．10 日前からの最高血流速度変化と 1 日の歩数の散布図を図 5 に示す．

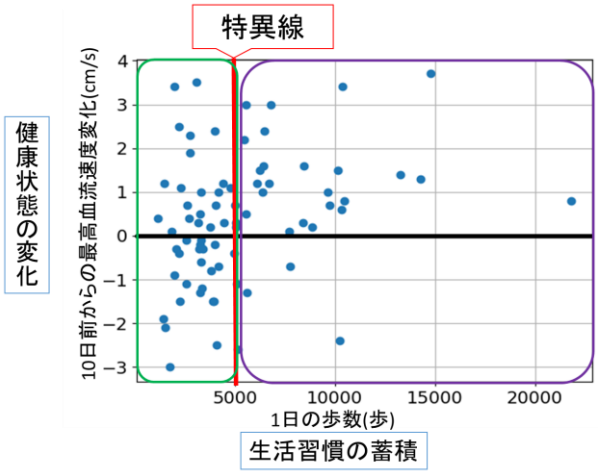

図 5　10 日前からの最高血流速度変化と 1 日の歩数の散布図

3.1 節で述べたように健康データのプラス変化，マイナス変化に着目して散布図を見ると，1 日の歩数が 5000 歩以下の範囲では 10 日間の最高血流速度のプラス変化，マイナス変化で同程度のデータのばらつきがみられるのに対し，1 日の歩数が 5000 歩より多い範囲ではプラスの変化に多くのデータが偏っていることが分かる．よって，この散布図では，1 日の歩数が 5000 歩付近で特異線が存在すると言える．

3.3. 特異線が存在する確率の算出方法

説明変数の選択に用いた特異線が存在する確率の求め方について述べる．特異線が存在する確率の算出は，散布図を 4 つの領域に分割し，それぞれの領域の条件付き確率を求める，条件付確率の差の絶対値によって特異線が存在する確率を求める．また，散布図の分割線を変え，1 つのデータセットにつき複数の確率を計算し，最大値をそのデータセットにおける特異線が存在する確率として採用する．図 6 に示す散布図の分割モデルを例に詳述する．

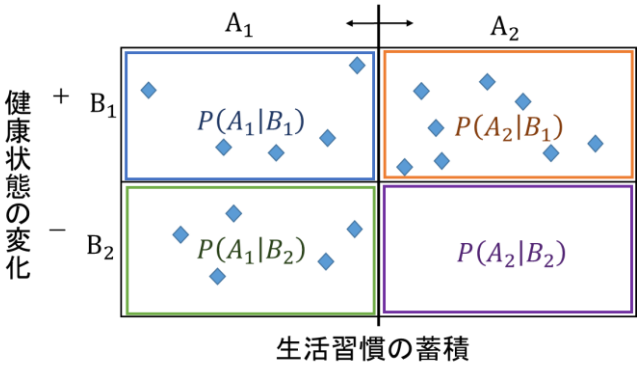

図 6　散布図の分割モデル

1）　散布図を 4 つの領域に分割

散布図は健康データのプラス変化（B_1）およびマイナス変化（B_2），生活習慣データの蓄積をA_1，A_2に 2 分する線によって 4 つの領域に分割する．

2）　条件付確率の算出

生活習慣データの蓄積の範囲の条件下（A_1もしくはA_2）におけるB_1およびB_2の条件付き確率を求める．条件付確率は式 (1) によって求める．

$$P(B|A_i) = \frac{P(A_i \cap B_j)}{P(A_i)} \qquad (1)$$

3）生活習慣データの範囲別に偏りの算出

生活習慣データの各範囲で健康データ

のプラスの変化とマイナスの変化に含まれる確率に差（偏り）があるかを算出する．

4） 散布図全体の偏りの算出

この 2 つの範囲におけるデータ数の偏りの差が大きいほど散布図全体のデータ分布に特異的な偏りがあると言える．つまり，特異線が存在する確率が高いと言える．特異線が存在する確立は式（2）によって求められ，0 から 1 の値をとる．

$$P_{sl} = \left| |P(B_1|A_1) - P(B_1|A_2)| - |P(B_2|A_1) - P(B_2|A_2)| \right| \quad (2)$$

4. 実験および評価

提案手法の有用性を示すために，従来手法との比較検討を行う．本実験の処理フローを図 7 に示す．本実験において，手動でデータ解析を行うための DB クライアントである「Clementine」（SPSS 社）を用いた．

本学の大学生 12 名（男性 7 名，女性 5 名）を対象とし，データの取得期間は 6 月 1 日～8 月 31 日の 3 ヶ月間とした．各学生が関心のある健康データ項目を決定し，健康状態の変動に関連のありそうな生活習慣について測定および記録を行った．

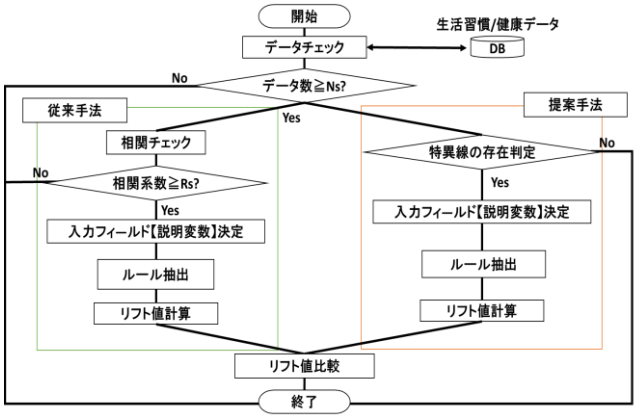

図 7　本実験の処理フロー

4.1. 説明変数の選択

提案手法では，コンピュータが自動的に算出した特異線が存在する確率が 0.5 以上のものを説明変数として選択した．従来手法による説明変数の選択は，自動健康データマイニングの手法に準じて手動で行った．本実験では，R_sの閾値を 0.300 とした．

その結果，全学生合わせて提案手法では 35 件，従来手法では 19 件の生活習慣が説明変数として選択された．提案手法では，相関係数が低いために従来手法では選択されることのなかった生活習慣が 16 件選択された．また，従来手法と提案手法で蓄積期間または遅延期間が異なる説明変数は 15 件となった．

4.2. 決定木分析およびルール抽出

提案手法および提案手法によって選択された説明変数を基に，学生ごとに決定木分析を行った．

決定木とは，単純なルールの連鎖を使って，大量のレコードをより少量のレコードの集合に逐次分解していくために用いられる仕組みである．一般的に使われているアルゴリズムは，CART と CHAID である．本実験では比較的新しいアルゴリズムである C5.0 を用いた．C5.0 は，各ノードを 1 つの独立変数を用いて複数のノードに分割させて決定木を生成する．過学習（分類に用いるデータ特有の歪みに適合しすぎた状態）させてから枝刈りを行い，正確な予測モデルを構築する[6]．生活習慣を入力変数（説明変数），健康状態を出力変数（目的変数）として決定木を生成した．

生成された決定木から，各ノードにおける目的変数のシンボル値の割合と確信度（Confidence）を基にルールを抽出する．確信度とは，条件Xがどれだけ事象Yに含まれるかを示している[7]．本研究では，確信度が 0.2 以上で目的変数のシンボル値「高い」または「低い」が含まれる割合の最も高いノードを選択し，ルール抽出を行った．

その結果，提案手法では 12 名中 11 名，従来手法では，12 名中 7 名のデータからルールを抽出することができた．

4.3. リフト値

決定木分析により抽出されたルールの評価にはリフト値を用いる．リフト値とは，主にマーケットバスケット分析における指標の一つとして使われており，条件 X のときに事象 Y がどの程度起こりやすいかを示す指標である．リフト値が 1.00 以上でそのルールが有効であるといえ，値が大きいほど良いルールであるといえる[7]．各ルールにおけるリフト値を表 1 に示す．

学生 12 名中 9 名において提案手法のリフト値が従来手法のリフト値よりも高い結果が得られた．

5. 結論

特異線の存在によって説明変数を選択する提案手法では，従来手法では選択されなかった説明変数が 16 件選択された．つまり，相関係数は低いが，特異線を持つような特徴をもった散布図になっていることが確認できた．よって，特異線が存在する良いルール抽出に適した説明変数であるものの，従来手法では相関係数が低いために選択されなかった説明変数を選択する

表 1　各ルールにおけるリフト値

	提案手法	従来手法		提案手法	従来手法
学生 A	1.71	—	学生 G	1.57	—
学生 B	1.48	—	学生 H	2.39	2.20
学生 C	2.22	2.50	学生 I	2.57	1.10
学生 D	1.88	—	学生 J	2.23	—
学生 E	1.83	1.77	学生 K	1.85	1.50
学生 F	—	2.70	学生 L	1.63	1.63

ことができたと考えられる.

抽出されたルールのリフト値を求めた結果, 学生 12 名中 9 名が従来手法よりも提案手法の方が高い値を得ることができたことから, 相関係数によらず説明変数として適切な生活習慣データが存在することが確認できた.

以上より, 相関係数によらず特異線に基づき説明変数を選択する提案手法は, 従来手法で選択できなかったルール抽出に有効な説明変数を選択することが可能であり, より的確な健康管理ルールを示すことができることから, 有用であると考えられる.

参　考　文　献

[1] 厚生労働省, “”平成 28 年版厚生労働白書－人口高齢化を乗り越える社会モデルを考える－”, 2016.

[2] 厚生労働省, ”平成 27 年度　国民医療費の概況”, 2017.

[3] 竹内裕之, 児玉直樹, 高橋慎吾, ”個人の体脂肪率と生活習慣との相関ルール生成にデータの季節変動が 及ぼす影響”, DEIM Forum 2015, G5-2.

[4] H. Takeuchi and N. Kodama, “Validity of Association Rules Extracted by Healthcare-Data-Mining”, Proc. 36th Annual International Conference of the IEEE EMBS, pp.4960-4963, 2014.

[5] 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, E1-5.

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

[7] 亀井靖高, 森崎修司, 門田暁人, 松本健一, ”相関ルール分析とロジスティック回帰分析を組み合わせた Fault-prone モジュール判別方法”, 情報処理学会論文誌 Vol.49 No.12, pp.3954-3966, 2008.