健康診断データからの時間的変化パターンの相関抽出

4W－10

白石 將，田中 秀俊
三菱電機 (株) 情報技術総合研究所

1. はじめに

健康診断データを用いて健康を害しつつある人間を発見するためには，不健康状態の前兆を示す知識が必要である．そのために，データベースから相関ルールを抽出するデータマイニングシステム Knodias を開発し [1]，1回分の健康診断データから健康診断の調査項目間の相関ルール抽出を試みて，生活習慣と体の不調の関連を示すような相関ルールを得た [2]．しかし，1回分の健康診断データに基づいた解析では，事象相互の静的な相関ルールの導出は可能でも，ある事象の生起に引き続いて別の事象が起こる，というような時間の概念を含む相関ルールを導くことはできない．そこで，連続した年に関する2回分の健康診断データを対象として，時間の概念を含む相関ルールの抽出を試みた．本稿ではその方法および結果について述べる．

2. Knodias の概要

Knodias による相関ルール抽出プロセスは，データの加工を行いマイニングに適した形式にする前処理，データから相関ルールを抽出するマイニング処理，得られた相関ルールから不要相関ルールを削除する後処理の3段階からなる．

相関ルール抽出のマイニングアルゴリズムは，アイテムの集合からなる不定長のレコードが複数集まって構成される形式 (以下「レシート形式」と呼ぶ) のデータベースを入力として処理を行う．相関ルールとは「A → B」の形式をしたルールであり，「データベース中で A と B を同時に含むレコードが多い」ことを意味する．ここで A と B はアイテムの集合を表し，A を条件部，B を結論部と呼ぶ．例えば時間的な関係を表す相関ルールの場合，時間的に先であるアイテムが条件部に，後であるアイテムが結論部に位置しなければならない．

以下，Knodias における前処理 / マイニング処理 / 後処理の内容について簡単に述べる．

前処理 解析対象のデータベースに対し，連続値属性の離散化 / 不要属性 (値) の削除 / 指定条件を満たすレコードの抽出等の加工を行う．さらに，各属性値名に対して属性名を結合してアイテムに変換することにより，レシート形式データを生成 / 出力する．

マイニング処理 前処理が出力したレシート形式データにおけるアイテムの同時出現の頻度を数え上げていくことにより，出現に相関のあるアイテムを探索する．出現頻度の期待値を考慮し，アイテムの組を作ったらその χ^2 値を算出し，自由度1の χ^2 分布を考えて，独立性の仮説が大きく棄却された組を相関ルールとして記録する [3]．処理の簡略化のため，抽出される相関ルールの形式は，結論部のアイテム数が1個であるものに限定している．また，相関ルール抽出の際，各アイテムに対して，相関ルールの条件部 / 結論部への出現の可否を設定可能としている．

後処理 マイニング処理が出力した相関ルールのうち，不要と判断されるものを削除後，残った相関ルールを表示する．要 / 不要の判断基準は以下の通り．ある相関ルールに対して，条件部にさらにアイテムを付加して生成された相関ルールを考える．アイテムを付加して制限を強くしたのにもかかわらず、相関の強さを示す指標である χ^2 値が増加しなかった場合，生成された相関ルールの χ^2 値は，もとの相関ルールの χ^2 値を反映しているに過ぎないと判断して削除する [4]．

3. 解析

解析対象のデータは，社内の 96/97 年の健康診断データであり，問診 / 身体測定 / 検査などの属性に関するデータが記録されている．解析作業は，96/97 年のデータを統合した後，統合後のデータを Knodias で処理することにより実施した．

解析の目的は，不健康状態の前兆を示す知識を得ることである．そこで，不健康状態の発生が「96 年から 97 年にかけてのデータの変化 (自覚症状に関する問診項目や検査項目などの値の変化を想定)」，また前兆が「96 年のデータおよび，96 年から 97 年にかけてのデータの変化 (生活習慣に関する問診項目などの値の変化を想定)」で表現されると考えることにした．従って，96/97 年のデータより，各人間の 96 年のデータおよび，96 年から 97 年にかけてのデータの変化を抽出して1つのレコードにまとめることにより，データの統合を行った．ここで，96 年から 97 年にかけての変化は，以下のように求めた．

- 連続値属性に関しては，97 年の属性値から 96 年の属性値を引いた値をデータ変化の属性値とする．
- それ以外の属性に関しては，96/97 年の属性値を表す文字列を，文字列「から」を挟んで連結し，文字列「へ」を最後に連結したものをデータ変化の属性値とする．例えば，問診項目に関する属性に対し，96 年の属性値が「いいえ」，97 年の属性値が「はい」であった場合，変化を表す属性値は「いいえ から はい へ」になる．なお，96/97 年の属性値が等しい場合には，データの変化は算出しない．

但し，年齢や性別などの属性に関しては，変化を求めることには意味がないので，変化の算出は行わない．

また，解析作業の便宜を図るため，属性名は，それが

Extracting Association Rules of Temporal Patterns from Medical Checkup Database
Masashi SHIRAISHI, Hidetoshi TANAKA
Mitsubishi Electric Corporation

96年のデータの場合には文字列「(96)」を，また変化を表すデータの場合には文字列「(96から97への変化)」を付加したものを用いた．

以上に述べた方式を健康診断データに適用することにより，結果的に，488属性(うち96年のデータに対応する属性数:248，96年から97年にかけてのデータの変化に対応する属性数:240)，2,872人分のデータが作成された．

データ統合後のKnodiasによる処理内容を以下に示す．

前処理 連続値属性の離散化後，頻度が300以上であるような属性値を一律に削除した．これにより，不健康状態に関連するような頻度の低い属性値に，特に注目することができる．さらに，データをレシート形式に変換して出力した．

マイニング処理 レシート形式において，頻度が10以上のアイテムを対象として，96年の属性値に対応するアイテムは結論部に来ないように指定を行ってマイニングアルゴリズムを適用した．χ^2値が10以上である相関ルールを抽出した結果，114,475個の相関ルールが得られた．

後処理 前述の，χ^2値を基準とした不要相関ルール削除を実施した結果，25,396個の相関ルールが残った．

4. 相関ルールの調査結果

4.1. 冗長な相関ルール

得られた相関ルールを調査した結果，1回分のデータより得られる相関ルールから導出可能な，冗長であると判断できるものが非常に多かった．以下にその例を示す．

- 相関ルール『(96) 目が痛い: いつも→(96から97への変化) 目が痛い：いつも から 時々 へ』．条件部が結論部を包含していることから，冗長であることは明らかである．
- 相関ルール『(96) 目が痛い: いつも→(96から97への変化) 目が疲れる：いつも から 時々 へ』．1回分のデータより『目が痛い: いつも→目が疲れる：いつも』なる相関ルールが得られ，また「(96) 目が疲れる：いつも」は「(96から97への変化) 目が疲れる：いつも から 時々 へ」を含む．従って，もとの相関ルールは冗長であると判断できる．
- 相関ルール『(96から97への変化) 目が痛い: いつも から 時々 へ→(96から97への変化) 目が疲れる：いつも から 時々 へ』．1回分のデータより『目が痛い: いつも→目が疲れる：いつも』および『目が痛い: 時々→目が疲れる：時々』なる相関ルールが得られる．従って，もとの相関ルールを96年と97年に分解して考えることにより，冗長であると判断できる．

そこで，Knodiasが出力する相関ルールのセットに対し，上記のような冗長な相関ルールを自動的に削除する処理を行った．この処理では，結論部のアイテムと，条件部を構成する各アイテムとを，上述のような視点に基づいて比較することにより，その相関ルールが冗長か否かを判定する．アイテム間の比較においては，96年の1回分のデータより，頻度が10以上のアイテムを対象として抽出されるχ^2値が10以上であるような相関ルール(37,982個)を利用した．冗長相関ルール削除処理の結果，5,549個の相関ルールが残った．

残った相関ルールを調査したところ，まだ冗長な相関ルールが多く含まれていた．これは，冗長相関ルール判定の際の基準が甘いためであり，さらなる工夫が必要であると考えている．

さらに上記処理方法では，連続値属性に関する冗長相関ルール，例えば『(96から97への変化) ヘマトクリット:2.8超→(96から97への変化) 赤血球数：37超』(1回分のデータから「ヘマトクリットと赤血球数には強い相関がある」ことが別途導出されるので，冗長であると判断される)が削除できない．従って，そのような冗長相関ルールの削除の仕組みも実装する必要がある．

4.2. 有用である可能性がある相関ルール

前述のように冗長な相関ルールが多いが，有用である可能性がある相関ルールも発見された．例えば，不健康状態の発生を表すと考えられるアイテム「(96から97への変化) 固いものが噛めない: いいえ から はい へ」に着目したところ，以下のような相関ルールが発見された．

- 『(96) 手足にむくみがある: 時々，(96) 下肢が痛い: 時々，(96) 手指の動きが悪くなる: 時々→(96から97への変化) 固いものが噛めない: いいえ から はい へ』

一方，96年分のデータのみを用いて，上記相関ルールの条件部と，アイテム「(96) 固いものが噛めない: いいえ」との相関を調査したところ，これらの事象の間には負の相関がある，つまり，これらの事象が同時に生起する頻度は少ない，ということが確認された．従って，上記相関ルールの条件部に相当する症状が，結論部に相当する症状発生の前兆である可能性があることがわかる．この相関ルールが正しければ，「手や足が痛いなどの不具合がある人は，しばらくして固いものが噛めないという症状も発生することが多い」という知識が得られることになる．

5. おわりに

連続した年に関する2回分の健康診断データを対象として，時間を含む相関ルールの抽出を試み，ある症状の前兆を示している可能性がある相関ルールを発見することができた．今後は，冗長相関ルール削除の方式および，連続値属性の処理方式に関する検討を進める予定である．

参考文献

[1] 白石他: データマイニングシステムKnodiasの構成, 第56回情報処理学会全国大会 2W-5 (1998).

[2] 田中他: データマイニングシステムKnodiasによる健康診断データの解析, 第56回情報処理学会全国大会 2W-9 (1998).

[3] 三石他: Knodiasにおけるデータマイニング方式, 第56回情報処理学会全国大会 2W-6 (1998).

[4] 川上他: 相関係数による数理的フィルタリング機能検証, 第10回データ工学ワークショップ[DEWS'99] (1999).