2W－9

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

田中 秀俊[1], 白石 將[1], 小幡 康[1], 山崎 高日子[1], 三石 彰純[1], 加藤俊夫[2], 奥田武正[2]

三菱電機(株) 情報総研[1], 系統変電交通システム事業所健康増進センター[2]

1 目的

健康診断データをKnodiasを用いて解析した。目的は生活習慣病予防のための手がかりの探索、健康診断データにおける知識発見プロセスの明確化、Knodiasの評価の3つである。本稿ではこの中の知識発見プロセスの明確化、すなわち、健康診断データをデータマイニング技術でどのように解析できるかという例を挙げることを主目的とする。

健康診断データの解析では、多量の「正常な」人たちの中に埋もれた比較的少数の「やや健康を損ない始めた疑いのある」人たちの発見を試みることになる。他の類似応用としては、故障の原因分析などが挙げられる。

対象としたデータは、社内の健康診断データで、問診、身体測定、検査、体力測定などからなる、全561属性、5770人の表形式のデータベースである。本稿ではこれをKnodiasを用いて解析した結果と、その解析過程について報告する。

2 解析方法

健康診断データからの知識発見のプロセスは、前処理、相関発見、後処理を繰り返す構造を採用した。この各段階でどのような手続きを施したかを以下に列挙する。

2.1 前処理

表形式の健康診断データから、相関発見エンジンの処理に向いたレシート形式へと変換する。この際に、データを圧縮して解析を高速化するとともに結果の意味を整える目的で、主に以下の5つの手続きを施した。

離散化 －Knodiasは、離散化すべき属性を自動的に検出し、離散化の提案を行なう。さらに、血液検査など、正常値と異常値がわかっているようなケースでは、正常値の境界を手で入力して離散化提案をオーバライドする。本実験では90属性の自動提案があり、うち50属性をオーバライドした。

属性値のグルーピング －離散値属性において、いくつかの離散値をまとめてひとつの離散値にする。本実験では39属性に対してグルーピングを施した。

属性値削除 －ある属性値をnull値(無値)に置き換える処理を指す。本実験では、一律に300人以上の頻度を持つ属性値を全部削除する方法を採用している。これにより、健康相談を要することを示唆する属性値に、特に注目することができる。

属性削除 －属性値が一種類の219属性を削除した。

アイテム化 －属性値に属性の識別を付け加え、アイテムにする。本実験では、属性と値とを区切り符号を挟んで単に結合してアイテムにする方式を採用した。総アイテムは1973種類だった。

2.2 相関発見

表形式からレシート形式に変換された後に、レシート内の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(アイテムセット)を作ったらそのχ^2値を算出し、相互依存が否定できるかを検定する。

2.3 後処理

相関発見手法により、大量の相関ルールが生成される。その中から本当に有用な面白いルールを抽出する作業を支援するために、Knodiasにはソート、特定アイテムによる検索、不要ルールの記録、特定ルールによる元データの検索などの機能を用意している。

3 結果

頻度10以上300未満のアイテム670個に関して、χ^2値が5以上のものを対象に相関発見アルゴリズムを適用したところ、得られた正の相関ルールは全部で74974個だった。ルールのアイテム数別の数の内訳と探索の累積所要時間は以下の通り。

これをすべて検討することは、専門知識を要することでもあり、今後に譲る。代わりに以下では、得られたルールから有用ルールを抽出するプロセスの例を示す。本報告では、アイテム数の多いルールと、アイテム2個のルールの調査を例にとった。

アイテム数	ルール数	所要時間(秒)
2	15976	5.9
3	20499	19.8
4	20075	29.2
5	12865	36.7
6	4770	41.9
7	749	45.9
8	40	49.7

(apricot LS550, pentium 200MHz, 主記憶80MB)

Health check analysis using data-mining system Knodias
H. Tanaka, M. Shiraishi, T. Yamazaki, A. Mitsuishi, T. Kato, T. Okuda
Mitsubishi Electric Corp.

アイテム8個のルール40個は実質的には5つの組合せが発見されていることを意味する。そのうち4つの組合せについては、頻度が10未満の9個組の存在を示唆していた。その例を以下に示す。

9個組: 頻度9人

頻度	アイテム
240	「異味を感じる」:「ときどき」
115	「皮膚が黄色く感じる」:「ときどき」
262	「手指の動きが悪くなる」:「ときどき」
239	「手指が痛い、はれる」:「ときどき」
231	「手指が冷える、白くなる」:「ときどき」
254	「下肢が痛い」:「ときどき」
221	「下肢がしびれる」:「ときどき」
110	「下肢がはれる」:「ときどき」
211	「歩きづらい」:「ときどき」

アイテムに注目して絞る例として、「パンにはジャムなどを厚くぬる」(頻度265)というアイテムに注目したところ、以下のような25のアイテムとの相関ルールが得られた。

組頻度	期待値	$\chi 2$	頻度	アイテム
37	13.7	43.6	299	塩辛いものを?: よく食べる
25	13.5	10.7	295	人に頼りすぎ、自主性に欠けると思う: いつも
27	13.5	15.1	293	目が疲れる、かすむ: いつも
25	12.4	13.9	271	朝食にめん類を?:1杯
25	12.3	14.4	268	夕食に魚、肉、大豆製品を?: たくさん
23	12.1	10.8	263	胸X所見1: 治癒型
27	11.8	21.7	256	漬物類は?: たくさん
22	11.3	11.0	247	TG:200-300
25	11.0	19.6	239	自分の性格が嫌: いつも
25	10.7	21.1	232	光りをみると虹が見える: ときどき
20	10.1	10.6	220	就業年数:35-45
24	10.0	21.3	218	喫煙年数:31.5-40.5
21	9.8	13.8	214	急に気力が低下、仕事能率が低下: いつも
20	9.7	11.9	211	歩きづらい: ときどき
21	9.2	16.3	201	一日の喫煙本数:25-35
20	9.2	13.6	201	胃X所見2: 胃炎
20	9.2	13.8	200	総合判定3: 高TG要観察
22	9.0	20.6	195	物忘れ: いつも
20	8.5	16.6	186	鼻が悪い: いつも
21	8.2	21.5	179	朝食に米飯を?:2杯
22	7.9	27.5	171	思うようにならないと怒る: いつも
20	6.9	26.5	151	何事もおっくう: いつも
21	6.2	37.4	136	甘い飲料を1日あたりに?:2本程度
26	5.9	72.9	129	料理に砂糖を使用しますか: たくさん使う
24	4.2	98.6	92	甘い菓子を1日に?: たくさん

4 考察

本実験の解析プロセスとしては以下の4種類を試みた。

- 取り扱うアイテムを内容に関わらず頻度範囲で限定
- アイテムの多いルールに注目し、関連ルールを調査
- 高頻度高 χ^2 値のルールのグルーピングを検討
- あるアイテムを含む2アイテムのルールを調査

多アイテムのルールの例では、相関発見を通じてあるクラスタが検出できている。アイテム9個のルールを満たす9人はどの時点でどのくらいに絞られているのかを調べるために、アイテム2個とアイテム3個のルールの一部を調査したところ、『ときどき「異味を感じる」「皮膚が黄色く感じる」「下肢がはれる」』という3個のアイテムで既に21人に絞られていることがわかった。多アイテムのルールで表現できるクラスタに、そのアイテムの部分集合の2アイテムや3アイテムのルールの集団は所属しつつあるという見方が可能と考えられる。

2アイテムの高頻度および高 χ^2 値のルールからは、意味的に冗長な部分が判明した。また人間の身体の性質上当然と思えるルールも多く示されていた。頻度や χ^2 値の高過ぎるルールは、自動的にアイテムグルーピングを行なった方がよいと言える。

1アイテムに注目した2アイテムのルールの例では、「パンにジャム等をいつも厚く塗る」を選択したところ、脂肪(TG)、喫煙本数、胃炎、砂糖使用などとが比較的高い相関があるという結果が見えた。これらを通じて、生活習慣と高脂肪や胃炎との関係が、人間の目には見える。これはたまたま人間の目で偶然発見したものに過ぎず、このような関係をどのように自動的に提示できるようにするかは、今後の重要な検討課題である。

5 知識発見プロセスの定式化(案)

結果および前章の考察から、健康診断データからの知識発見プロセスとして、以下のような形態が有効と思われる。これは、他の類似応用、例えば故障の原因分析などにも有効と思われる。

【初期調査】 離散化は正常値と異常値が事前にわかるもののみ人力設定、残りは自動提案・アイテム化自動提案のまま、アイテム頻度範囲を一律に3～5%未満に絞り、マイニングを行なう。後処理として、特にアイテムの多いルールに関してそこに含まれる2アイテムのルールを調査する。

【冗長性軽減反復】 高 χ^2 値のルール、例えば2500以上を目安として、アイテムグループとして登録、再び前処理し、アイテム頻度範囲も上下限をやや広げてマイニングを行なう。引き続いて、χ^2 値1000を目安に不要ルールに登録、もしくはアイテムグループに追加登録を行なって再び前処理から反復する。

【詳細探索反復】 離散化、属性値グルーピング、属性値削除を用いて、各属性において属性値を出現頻度3～5%、2～3個程度を目安にまとめあげ、代わりにアイテム頻度範囲の設定をせずにマイニングを行なう。

6 まとめ

健康診断データをデータマイニング実証システムで解析しその過程を報告することにより、データマイニング技術を用いた知識発見プロセスの一例を示した。今後の主な課題としては、解析結果の考察、アイテムグルーピングを行なう χ^2 値の基準値の決定の2点が挙げられる。

参考文献

[1] Agrawal,R. et al. Fast Algorithms for Mining Association Rules. *VLDB94*, (1994).

[2] Tanaka,H. et al. Database Reduction for Discovering Health-Care Rules. *PADD98*, submitted.

[3] 加藤他. データマイニングの手法を用いた検診データの解析. 第37回近畿産業衛生学会, (1997).