データマイニングシステム Knodias 簡易版の設計

1T－2

白石 將，田中 秀俊，野本 弘平
三菱電機 (株) 情報技術総合研究所

1 はじめに

大量のデータの中から規則性を発見するデータマイニングの一手法として相関ルールの抽出がある．

我々はこれまで相関ルール抽出用データマイニングシステム Knodias の開発を進めてきた [1]．現在，手軽にマイニングを行えるような環境を作ることを目的として，Knodias の前処理 / 後処理部分を表計算ソフトのマクロ言語で記述し，全体をその表計算ソフト上で動作できるようにした Knodias 簡易版の開発を進めている．

従来の Knodias[1] は，一般的な表形式 (次章で述べる「スプレッドシート形式」) を対象として相関ルール抽出を行っていた．しかし，実際のデータには，そのような整った形式のデータ以外にも，単に記録として残すためだけに生成されたデータなど，取り扱いにくい形式のデータも多い．これまで，そのようなデータを扱う際には，それぞれのデータ形式毎に，データを一般的な表形式に変換するためのプログラムを作成して変換を行うなどの処理を事前に行う必要があった．そのような労力を軽減するために，柔軟性の高いデータ形式である ST 形式を考案し，Knodias 簡易版で扱えるように設計を行ったので報告する．また，結果表示において，相関ルール数が多い場合でもユーザが楽に解析作業を行えるように，工夫を行ったので合わせて報告する．

なお，簡単のため，本稿内では Knodias 簡易版のことを，単に「Knodias」と表記することにする．

2 解析対象データ形式の拡大

相関ルール抽出のアルゴリズムは，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[1]) のデータベースを入力として処理を行う．各レコードには識別子 (以下「ID」と呼ぶ) が付与されており，同一 ID を持つレコードを単位 (以下「ID 単位」と呼ぶ) として相関ルール抽出が行われる．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 を含む ID 単位は同時に B も含むことが多い」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」，また B を「結論部」と呼ぶ．

一方，実際の代表的なデータ形式としては以下のものが考えられ，Knodias の前処理では，これらの形式を入力としてレシート形式に変換できることが望ましい．

スプレッドシート形式 各レコードが (ID, 属性 1 に対応する属性値, 属性 2 に対応する属性値, …) という並びで構成される表形式 (表 1参照)．属性の並びは別途定義されている．

表 1: スプレッドシート形式

	属性 1	属性 2	属性 3	属性 4	属性 5
ID1	値 11	値 12	値 13	値 14	値 15
ID2	値 21	値 22	値 23	値 24	値 25
ID3	値 31	値 32	値 33	値 34	値 35
⋮	⋮	⋮	⋮	⋮	⋮

表 2: 単純トランザクション形式

ID1	属性 1	値 11
ID1	属性 2	値 12
ID1	属性 3	値 13
⋮	⋮	⋮
ID2	属性 1	値 21
ID2	属性 2	値 22
ID2	属性 3	値 23
⋮	⋮	⋮

単純トランザクション形式 各レコードが (ID, 属性, 値) の 3 つ組から構成される表形式 (表 2参照)．ID や属性は複数列の直積で与えられていても良いが，値は 1 列とする．

我々はスプレッドシート形式と単純トランザクション形式とを包含するような，より一般的なデータ形式である「Spreadsheet-in-Transaction 形式 (以下「ST 形式」と記述する)」を考案し，Knodias の前処理において ST 形式からレシート形式への変換を行うように設計を行った．本設計は，様々なデータ形式に対する前処理の労力を軽減することを狙いとしている．

以下，ST 形式について表 3に基づいて説明する．値の列は単純トランザクション形式では 1 列だったが，ST 形式では複数列から構成されるように拡張されている．この時，属性は複数の値の並びを束ねるものに過ぎなくなり，それぞれの値の意味 (普通のスプレッドシート形式における属性に相当するもの．本稿では区別のた

表 3: ST 形式

ID1	属性 1	値 111	値 112	値 113
ID1	属性 2	値 121	値 122	
ID1	属性 3	値 131	値 132	値 133
⋮	⋮	⋮	⋮	⋮
ID2	属性 1	値 211	値 212	値 213
ID2	属性 2	値 221	値 222	
ID2	属性 3	値 231	値 232	値 233
⋮	⋮	⋮	⋮	⋮

Design of Portable Data Mining System "Knodias"
Masashi SHIRAISHI, Hidetoshi TANAKA , Kohei NOMOTO
Mitsubishi Electric Corporation

[1]本稿におけるデータ形式の名称は一般的なものではなく，あくまでも本稿内でのみ通用する名称である．

表 4: 属性名マスター

	値の１列目	値の２列目	値の３列目
属性１	サブ属性	サブ属性	サブ属性
属性２	サブ属性	サブ属性	
属性３	サブ属性	サブ属性	サブ属性
⋮	⋮	⋮	⋮

め「サブ属性」と呼ぶことにする) は，レコードにおける属性および値の位置によって定まることになる．この，属性および値の位置と，サブ属性との対応を記述したテーブルを「属性名マスター」と呼ぶことにする．属性名マスターの形式を表 4に示す．

病院の診療記録データベースがST 形式である場合について説明する．ID 欄には個々の患者名が記述されている．また属性欄には診療科名，例えば [内科] や [循環器科] が記述されている．そして，値の欄には診療科毎の検査結果が記述されている．例えば属性 [内科] を含むレコードには [尿潜血] や [脂肪厚] などの検査結果が，また属性 [循環器科] を含むレコードには [脈拍] や [血沈] などの検査結果が記述されている．これら [尿潜血][脂肪厚][尿潜血][脂肪厚] などが，前述のサブ属性に相当する．この時，属性名マスターには，「属性 [内科] の値の１列目のサブ属性が [尿潜血]，２列目のサブ属性が [脂肪厚]」などの対応が表の形式で記述されている．

以上のように，ST 形式は単純トランザクション形式の値列を複数持つようにしたものであり，これを包含する．また，スプレッドシート形式は，ST 形式において属性列を指定しない特殊な場合に相当する．従って，ST 形式からレシート形式への変換機能を用意しておけば，単純トランザクション形式やスプレッドシート形式からレシート形式への変換は可能になる．また，表形式にもなっていないような未整理データからの相関ルール抽出が必要な場合でも，スプレッドシート形式や単純トランザクション形式まで整理する必要はなく，ST 形式まで整理できれば良い，ということになる．

以下，Knodias の前処理における，ST 形式からレシート形式への変換の流れを示す．

1. ST 形式において，各列に関してユーザに ID 列 / 属性列 / 値列の指定をさせる．ID 列に関しては，複数の列を指定した場合，その直積が ID となる．属性列に関しても同様である．
2. 列指定が完了すると属性名マスターの形が決まるので，Knodias は属性名マスターを自動生成し，ユーザにサブ属性名の入力をさせる．
3. サブ属性毎に値の頻度分布を調査して，その結果をユーザに提示する．あるサブ属性に関する頻度分布を得るには，まずそのサブ属性に対応する属性を含むレコードのみを ST 形式より取り出して，さらにそのサブ属性に対応する列を取り出してその値に関してソートし，調査集計すれば良い．
4. 頻度分布をもとにして，ユーザに連続値の離散化などのデータ編集作業をさせる．
5. データ編集後，ST 形式を ID に関してソート，さらに同一 ID 部分をまとめて１つのレコードとして，値をアイテムに変換する．これがレシート形式である．値からアイテムへの変換は，サブ属性名と値とを結合することにより行う．

3 相関ルール表示の工夫

相関ルール抽出エンジンにより出力される相関ルールは，対象とするデータの内容や相関ルール抽出の際のパラメータの設定によっては膨大なものになることがある．そのような場合，相関ルールの表示ができなくなったり，表示ができる場合でも長時間を要する，という事態が生じる上，人手による相関ルールの解析作業も困難になる．この問題を解決するため，Knodias を，出力された相関ルール数をその結論部 / 条件部に含まれるアイテム毎，また構成アイテム数[2]毎に集計した表 (「ルール辞書」と呼ぶ) を作成するように設計した．ルール辞書の形式を表 5に示す．

表 5: ルール辞書

アイテム	構成アイテム数					
	2		3		4	
アイテム１	150	150	40	80	2	6
アイテム２	20	20	(5)	(9)		
アイテム３	10	10				
アイテム４	25	25	10	20	3	4
⋮	⋮	⋮	⋮	⋮	⋮	⋮

例えば，表5中で ○ で囲んだ部分はそれぞれ以下のことを表している．

- アイテム２を結論部に含み，３個のアイテムから構成されているような相関ルール数は５個．
- アイテム２を条件部に含み，３個のアイテムから構成されているような相関ルール数は９個．

ユーザがルール辞書の中で興味がある部分のセルを選択して表示命令を出すことにより，Knodias は該当する相関ルールを表示する．以上のような方式をとることにより，相関ルール数が多い場合でも，ユーザが興味のある範囲の相関ルールのみを簡便なインターフェースで見ることができる．

4 最後に

相関ルール抽出の解析対象とするデータ形式の拡大および，相関ルール表示の際の工夫について説明した．今後は，様々な形式の実データに対して Knodias を適用し，事前に行わなければならない前処理がどの程度軽減されたか，また，どの程度使いやすくなったか，等についての評価を行っていく予定である．

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 回情処全国大会 2W-5, 1998.

[2]相関ルールを構成するアイテムの数.