データマイニングシステム Knodias の構成

2W－5　白石 將，田中 秀俊，三石 彰純，山崎 高日子，小幡 康
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 (association rule) の抽出がある。有用な相関ルール抽出を効率よく行うために、マイニングアルゴリズム適用処理 (以下、マイニング処理と呼ぶ) を行う部分の他に、前処理部と後処理部とを備えたデータマイニングシステム Knodias(KNOwledge DIscovery Assistance System) を開発したので、その構成について報告する。

2　問題点とその解決策

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 を含むレコードは同時に 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、B を結論部と呼ぶ。

表形式のデータベースをマイニングする際に、各属性値をそのままアイテムと見做して処理を行うと、出力される相関ルールは意味がわかりにくいものになってしまう。従って、前処理においては各属性値に属性の識別を付与してアイテムとするアイテム化手続きが必須である。アイテム化手続きを施すことにより、例えば属性「身長」の属性値「178cm」は、アイテム「身長:178cm」に変換される。

さて、解析対象の表形式データベースにアイテム化手続きを施して出力されるレシートに直接マイニング処理を適用した場合、以下のような問題点がある。

A: アイテム数が膨大である場合、実行に長時間を要する。

B: ユーザにとって不要な相関ルールが多く生成される。不要な相関ルールとしては、ユーザの興味の対象外であるようなアイテムを含む相関ルール、もともと成立することが明白であり新規性のない相関ルールなどがある。

C: 連続値で表現される属性値に対するアイテムなど、情報が詳細である場合、一般にそれらのアイテムがデータベース中に出現する頻度は小さくなるが、マイニングアルゴリズムはあらかじめ定めた閾値よりも出現頻度が低いアイテムを除去するため、それらのアイテムを含むような相関ルールは生成されない。

D: 連続値で表現される属性値に対するアイテムなど，情報が詳細過ぎる場合や、アイテムの名前がコード化されている場合など，生成される相関ルールの意味付けが難しく、活用しにくいことがある。

以上の問題点を解決するために、Knodias における前処理 / マイニング処理 / 後処理では以下のような機能を持つ。

前処理　解析対象の表形式データベースに各種手続きを施して、マイニング処理の入力であるレシートを作成する処理。アイテム化の他に適用可能な手続きは以下の通り。

- 指定された属性値を null 値 (無値) に置き換える属性値削除手続き。問題点 A/B の解決に寄与。
- 不要な属性を削除する属性削除手続き。問題点 A/B の解決に寄与。
- 属性や属性値の名前を違う名前に変更する読替え手続き。問題点 D の解決に寄与。
- 連続値属性を適切な数の離散値の属性に変換する離散化手続き。問題点 C/D の解決に寄与。
- 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属性値グルーピング手続き。問題点 C/D の解決に寄与。
- アイテム化手続き適用後、いくつかのアイテムをまとめてひとつのアイテムにするアイテムグルーピング手続き。問題点 C/D の解決に寄与。

また、上記の手続きの他に、注目する属性の属性値を指定してその属性値を含むレコードのみを抽出する絞り込み (いわゆるセレクション) 手続きも適用可能とする。

マイニング処理　レシートにマイニングアルゴリズムを適用して相関ルールを抽出する処理。アルゴリズムは以下の特徴を持つ。

- ユーザは事前に各アイテムに対し、相関ルールの条件部や結論部がそのアイテムを含むことの可否を設定する。これを条件部結論部設定と呼ぶ。アルゴリズムは、設定に合う相関ルールのみを抽出する。問題点 B の解決に寄与。また、相関ルールの探索範囲が狭まるので問題点 A の解決にも寄与する。

後処理　マイニング処理により抽出された相関ルールの表示を行う処理。表示前に、以下の手続きを実行する。

- 相関ルールの中で、以前にユーザにより不要であると判断されたルールと同一であるものを除去する。問題点 B の解決に寄与。

Knodias: Data Mining System
Masashi SHIRAISHI, Hidetoshi TANAKA, Akitoshi MITSUISHI, Takahiko YAMAZAKI, Yasushi OBATA
Mitsubishi Electric Corporation

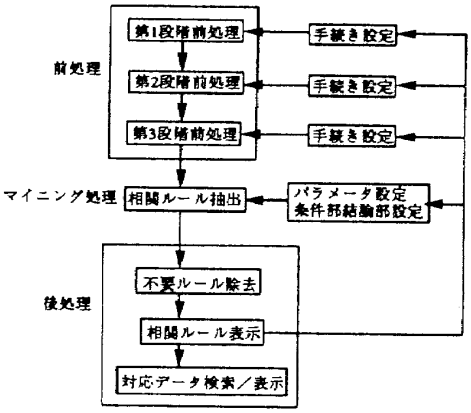

図 1: Knodias による処理の流れ

3 システム構成

「前処理→マイニング処理→後処理」の過程を一度だけ実行して全ての有用な相関ルールを抽出できることは稀であり、各処理の条件を変えて何度も試行錯誤を繰り返さなければならないことが多い。そこで、Knodias は図1のような処理の流れをサポートする。以下、図1を基にして Knodias の構成について説明する。なお、マイニング処理部に関しては [1][2][3] を参照のこと。

3.1 前処理部

第1段階前処理では離散化 / 読替え手続きの適用を、第2段階前処理では属性削除 / 属性値削除 / 属性値グルーピング / 絞り込み手続きの適用を、第3段階前処理ではアイテム化 / アイテムグルーピング手続きの適用を、それぞれ行う。

各段階の前処理の出力は、中間ファイルとして保存することができる。従って、データマイニングを行う時に毎回第一段階前処理より始める必要はなく、存在する中間ファイルから処理を開始することが可能である。

各段階の前処理に対する手続き設定では、適用する手続きを記録する手続きファイルの作成を行う。

前処理の実行は、手続きファイルの内容を手続き解釈ルーチンが解釈して、それに基づいて前段の前処理出力 (第1段階前処理の場合はもとの表形式データベース) を変換することにより行われる。各段階の手続き解釈ルーチンは同一であり、どの手続きファイルをどの段階で適用するかは、Knodias の GUI がガイドする形式で制限している。

各種手続きの中で連続値属性に対する離散化手続きとアイテム化手続きは必須であるので、これらの手続きは Knodias が自動的に作成し、作成された手続きが不満足な場合、ユーザが新たな手続きを作成してオーバーライドする、という方法を取る。離散化手続きの自動作成に際しては、属性値の集合より離散化すべき属性が判断され、離散化すべきであると判断された属性に対しては、ユーザが事前に指定した分割方法に基づく離散化手続きが作成される。

3.2 後処理部

後処理ではマイニング処理により得られた相関ルールを表示するが、ユーザはその際に不要な相関ルールを選択する。選択された相関ルールは不要ルールファイル中に保存され、以後のマイニング処理後は、不要ルールファイル中のルールは自動的に除去された上で表示されるようになる。

また、相関ルールの調査の便宜を考え、以下の機能を実装している。

- 特定のアイテムを含む相関ルールのみを表示する機能。
- もとの表形式データベース中で、指定された相関ルールに対応するレコードを検索・表示する機能。

3.3 GUI

前処理の手続き設定やマイニング処理の条件部結論部設定の際には、属性や属性値の関係が木構造で画面表示される。ユーザは画面内の属性ノードや属性値ノードを選択することで容易にこれらの設定を行えるようになっている。

表示用木構造は、属性階層とデータ辞書より生成される。属性階層とは、属性間の関係を木構造として表現したものであり、事前にユーザが別の GUI により作成しておく。また、データ辞書はもとの表形式データベースや各段階の前処理の出力に対応して作成されるファイルであり、そのデータ中に存在する属性名と属性情報を、属性ごとに記述している。属性情報には、属性値のリストや各属性値の頻度が含まれている。前処理の手続き設定やマイニング処理の条件部結論部設定の際には、前段階の前処理の出力に対応するデータ辞書に応じて、属性階層の各属性ノードに対する属性値ノードの付加や、存在しない属性ノードの除去などを行うことにより、表示用木構造が生成される。前処理の手続き設定の際には、さらに編集中の手続きファイルの内容も反映させた木構造が生成・表示される。

また、前処理の手続き設定の際に、データ辞書に含まれている属性値の頻度情報よりヒストグラムの生成・表示ができるようになっている。このヒストグラムを参照することにより、適切な手続き設定が可能になる。

4 最後に

現在、健康診断データベース等を対象として Knodias を評価している。今後は、その評価結果に基づき、前処理 / 後処理に必要な機能を再検討し、より使いやすいデータマイニングシステムを構築していきたいと考えている。

参考文献

[1] 三石, 他：Knodias におけるデータマイニング方式, 第56回情処全国大会 1998.

[2] 山崎, 他：Knodias におけるデータの性質に着目した相関ルール抽出の効率化, 第56回情処全国大会 1998.

[3] 小幡, 他：有限メモリ空間で相関ルールを抽出するマイニングアルゴリズム, 第56回情処全国大会 1998.