特　集　「共通データによる知識発見手法の比較と評価」

構造データ及び数値データに対する相関ルールマイニングの拡張

Extention of Association Rule Mining for Structured and Numerical Data

鷲尾 隆　大阪大学産業科学研究所
Takashi Washio　The Institute of Scientific and Industrial Research, Osaka University.
washio@ar.sanken.osaka-u.ac.jp, http://www.ar.sanken.osaka-u.ac.jp/washprjp.html

元田 浩　（同　　上）
Hiroshi Motoda　motoda@ar.sanken.osaka-u.ac.jp, http://www.ar.sanken.osaka-u.ac.jp/motoprjp.html

Keywords: association rule, graph structured data, frequent graph, numerical data, descretization.

1. はじめに

膨大なデータの中から有用な，あるいは興味のあるパターンを明示的な知識として発掘しようとするデータマイニングは，計算機能力を最大限に活用したアプローチである．データマイニングにおいてしばしば扱われる実問題データは，患者の病名とその症状のように各事例が目的クラスとその説明属性からなる表形式や，スーパーマーケットのレジ販売記録のようにアイテム（品物）の集合からなるトランザクション形式を有する．特にトランザクション形式のデータに関しては，効率よくアイテム間の相関ルールを導出するAprioriアルゴリズム [Agrawal 94] が提案されている．相関ルールマイニングは，計算効率の高さや得られる知識形式が簡素で理解しやすいなどの点が評価され，多くのデータマイニング手法の中でも多方面で実用に供されているものの1つである．しかしながら，この手法はトランザクションが単純なアイテム集合であることを前提としており，得られる相関ルールはトランザクションにおけるアイテムの共起関係を表すものに限られる．また，各アイテムは記号情報であることが前提とされており，実データにしばしば現れる個数や大きさなどに関する数値情報を扱う枠組みが用意されていない．本稿では，はじめに構造データと数値データに関する相関ルールマイニング手法の現状を概観する．次に相関ルールマイニングを行う際の実際的課題を吟味し，それに対して必要な手法拡張を検討する．更に，筆者等が行ってきた研究とその共通データへの適用結果を述べ，先の検討結果との比較を通じて今後に残された課題や展望を述べる．

2. 相関ルールマイニングの現状

多くの実問題では，トランザクションが単純なアイテム集合ではなく，何らかのデータ構造を有している．例えば，長袖シャツや半袖シャツというアイテムはいずれもシャツという一般的な上位カテゴリに分類されるように，トランザクション中の各アイテムが幾つかの上位カテゴリに分類される関係構造を有する場合がある．また顧客がある売場から選んだ品物の集合と他の売場から選んだ品物の集合が集まって1つのトランザクションを構成するように，アイテム集合の更に集合がトランザクションになる場合もある．前者の場合については，Agrawal等がAprioriを拡張してアイテムのTaxonomy構造付き相関ルールをマイニングする高速アルゴリズムを提案している [Srikant 97]．また後者についても，階層的アイテム集合構造を有するトランザクションに関し，やはりAprioriをベースとして多頻度パターンを高速に発見する手法が提案されている [Matsuzawa 00]．

より複雑な構造を有するトランザクションに関しても，近年，Aprioriアルゴリズムを拡張した種々の手法が提案されている．その1つは，トランザクション内において，アイテム間に空間や時間，あるいは優先順位などによる何らかの順序関係が存在する，即ち系列についてである．膨大な系列データからあらゆる特徴的パターンを抽出することを目指したデータマイニングの先駆的研究は，AgrawalとSrikantによるものであろう．これはアイテムからなるトランザクション間に系列順序の制約を課した上で，Aprioriに類似の原理を用いて多頻度の系列パターンを導こうとするものである [Agrawal 95]．また，記号系列データ中で頻繁に生起する半順序の記号連鎖パターンをエピソードと定義し，系列データから多頻度の

エピソードを抽出する手法も提案されている [Mannila 97b]. これらの手法は,系列データが有する順序ないしは半順序制約を巧みに用い,通常の Apriori と同程度かそれ以上の計算効率を実現している.日本国内でもハッシュ分割を用いた並列分散処理環境向きのアルゴリズムで,高速に系列パターンを導こうとする研究が行われている [Sintani 98]. ただし,これらの研究は何れも連続した記号ないしはトランザクション間の多頻度パターンを発見しようとするものであり,不定長個のワイルドカードが挿入されるような系列パターンを導くものではない.現状では,この問題に対する Apriori 拡張型の相関ルールマイニング手法は得られておらず,他の手法の枠組みでの研究が試みられている.例えば,PAC 学習理論の成果を用いて上限 k 文字以内の記号系列パターンの探索を低次多項式時間で行う高速アルゴリズムの研究 [Arimura 98] や,DNA シーケンスのアライメント解析や配列歩行問題に応用可能で,計算量,メモリ使用量の両面で高効率な探索手法の提案がなされている [Miura 98].

もう1つ代表的な複雑構造データはグラフである.グラフ構造パターンのマイニングは,特に化学分野で精力的に研究が行われてきた.代表的なマイニングシステムは CASE 及び MultiCASE と呼ばれるものである [Klopman 84, Klopman 92]. これらのシステムは化学分子構造の解析に特化された手法を採用しており,効率的に部分構造を発見できる反面,探索可能な部分構造は,分子内で枝分かれしていない1本の原子鎖に沿うものに限られる.より実際的に一般グラフ構造を有する特徴的部分構造を見いだす方法としては,分子構造データに前処理を施し,たとえばベンゼン環を有するか否かといった構造を特徴づける沢山の属性や命題に変換してから,C4.5 のような分類決定木や M5 などの回帰分類木,帰納論理プログラミングを適用するという方法も試みられている [King 96, Kramer 97]. しかし,この方法では前処理で人為的に取り上げられたパターンの組み合わせによる部分構造しか抽出できない.相関ルールマイニングの分野においては,Wang と Liu が木構造データ内に頻繁に現れる部分木構造をスキーマと呼び,そのようなスキーマを膨大な木構造データから効率的に発見する方法を提案した [Wang 97]. そして Web のリンク構造解析に適用した結果を報告している.ただし,この方法は閉ループを有するような一般グラフのパターン抽出には適用できない.以上は,特定のクラスや何らかの制限があるグラフ構造を扱う手法であるが,近年,より一般的なグラフ構造を扱う手法の研究も進んでいる. Graph-based Induction (GBI) は,グラフデータ内に頻繁に現れるノードペアを逐次チャンクし,チャンクの総体として多頻度部分グラフ構造を発見する手法である [Yoshida 95]. この手法は Greedy 探索を行うため,非常に少ない計算量でグラフ構造パターンを発見することができる.しかし,完全探索ではないので重要なパターンを見逃す可能性を排除できない.完全探索を試みる手法としては,Dehaspe 等が直接に化学物質の発癌性を特徴づける多頻度部分分子構造をマイニングする手法の提案を行った [Dehaspe 98]. これは帰納論理プログラミングを基本に用い,かつデータベースへのアクセス回数を最小化するように,命題や述語の探索において Apriori に類似した幅優先探索を使用し効率の向上を図っている [Mannila 97a]. この方法は従来の帰納論理プログラミングベースの手法よりも遙かに効率的であるが,それでも発癌性物質の多頻度部分構造探索の空間は膨大であり,実用時間内には高々6個の述語,即ち2,3個の原子からなる簡単な部分構造までの探索しかできなかった.

更に複雑な規則構造を有するトランザクションを扱う試みとして,一階述語論理でなければ表現困難な複雑で一般的な構造を有するデータの存在を前提とし,そのような構造を帰納論理プログラミング手法をベースとして,計算機学習させる研究が進められてきた [Dzeroski 96]. しかしながら,一般にこの手法は必要とされる計算量が膨大になる傾向があり,大量データに対するマイニングにはあまり適さない.

一方,Apriori をベースとする相関ルールマイニングは,記号アイテム間の共起相関を解析する手法であるため,数値で表される属性アイテムを含むデータに直接適用することは困難である.この場合には属性アイテムが取る値域を複数区間に離散化し記号ラベルに変換する必要がある.従来,機械学習においては,各属性アイテム値を χ^2-検定によって離散化するもの [Kerber 92],情報エントロピーを用いて離散化するもの [松本 93] など,多くの離散化手法が提案されている.しかしながら,その殆どは他の属性の影響を考慮せずに各属性個別に離散化を行っている.また,多くが ID3 に代表される分類木学習 [Quinlan 86] を対象としたものであるため,クラス情報の存在を前提としている.これに対し,相関ルールマイニングでは複数アイテム間の共起情報を得ることが目的であり,またクラス情報が存在しない.従って,従来の機械学習の枠組みにおける離散化はあまり適さない.このような観点から,クラス情報を必要とせずに複数属性アイテム値間の依存性に基づき,属性アイテム値空間を各単調領域毎に離散化する手法が提案されている [Yoda 97]. しかしながら,人間にとって理解容易な単調領域空間はせいぜい3次元までであり,それ以上の次元を持つデータへの適用は現実的でない.

3. 課題と望まれる手法拡張

構造を有するトランザクションは多種多様である.これに対しこれまでの研究の多くが,それらの内の代表的な幾つかの構造に特化した効率的アルゴリズムを探求して来ている.しかしながら,各種構造に特化した手法を1つ1つ積み上げるだけでは,データマイニングの現場

において既存手法に合致しない構造データを含む実問題に遭遇した場合に対処できない．一方，帰納論理プログラミングのように非常に一般的な構造を直接取り扱い可能な手法も提案されている．しなしながら，多くの実問題について必要計算量が膨大になり実用的な場合は限られるのが現状である．

科学技術や商業分野におけるデータマイニングニーズの殆どは，データベース内の膨大なトランザクションに含まれる物や事実の関係を整理することである．その際，具体的な物や事実が記述された関係提示の方が，データ解析者にとって理解しやすいことが多い．解析者はそのような具体的規則性の理解を積み上げながら，変数を含むような更にメタレベルの規則性を自ら納得しつつ発見することを好む．従って，計算機上のデータマイニング手法として，多くの場合は変数を含まない命題論理規則やそれに類する記述を得るものが妥当である．しかしながら一般に命題論理規則が表すデータ構造は，上述の Taxonomy やアイテム階層，系列などのデータ構造よりも複雑多様であり，一方で述語論理規則よりも遙かに単純である．また，データマイニングは計算機のパワーを最大限に動員して，なるべく多くの可能性の中から興味深い規則性を見つけ出したいというニーズを擁することが多く，その際にはなるべくなら与えた基準に合致するすべての相関ルールを導出することが望ましい．これらの点を鑑みると，命題論理規則で表される程度の複雑さを有する構造を，なるべく完全性を確保しながら実問題の規模に対して実用的な計算量の範囲内でマイニングし得る手法が望まれる．

一方，計算量理論や計算機学習理論の分野では，これまで計算量については問題の大きさに関して，何次の多項式オーダーないしは指数オーダーで解を見つけることが可能かが主要な論点であった．しかしながら Apriori はトランザクションサイズについて NP オーダーの計算量を必要とするにもかかわらず，データマイニングの分野で広く使われている．もちろん，アルゴリズムの複雑性は基本的速度を決定するため，重要であることには変わりはない．しかし，低速 2 次記憶装置へのアクセス数やメモリ管理のしやすさなど，実装上の高速性も手法の実用性を決める重要な要因である．従って，両方の観点を総合して実用時間内でマイニング可能な手法を開発する必要がある．

また先に述べたように，相関ルールマイニングにおける数値情報を含むデータの扱いは大きな課題として残されている．クラス情報を必要とせずに，複数アイテム間の共起情報を十分に保存しつつ，各属性アイテム値を離散化する手法が求められる．その際には属性アイテム値間の依存性を十分に反映した離散化が行われる必要がある．更に離散化粒度が細かすぎると，共起するアイテム記号の種類が増加する分，アイテム集合頻度が下がってしまい，本来得られるべき多頻度アイテム集合が導出できなくなってしまう．そのため，属性アイテム値間の依存性を表しつつ必要最低限の粗い離散化を行う必要がある．また，離散化操作は多数の属性アイテムを含むデータについても，解析者にとって十分理解容易な簡素な結果を提供せねばならない．

4. 構造データ・数値データに対する手法

以上のような検討の下で，筆者等は構造データや数値データを含むトランザクションに関する相関ルールマイニング手法の開発を行っている．以下に述べる手法は，まだ上記の課題のすべてを十分に解決するものではないが，現状可能な限りの挑戦を行っているものである．

4·1 構造データの扱い

現状では命題論理規則で表される構造を十分に効率よくマイニングする手法は知られていない．しかしながら，変数を含まない規則に書かれた各命題間の関係を，例えば論理関係を表すラベル付きリンクと命題を表すラベル付きノードとして表現すれば，一般グラフとして扱うことが可能である．そこで我々は，一般グラフの多頻度パターン及び相関ルールの全探索を高速に行う手法として，グラフの隣接行列表現と Apriori を組み合わせた Apriori-based Graph Mining (AGM) 手法を開発している [Inokuchi 99, 猪口 00]．

隣接行列の i 行及び i 列に相当するノードを第 i ノードとし，v_i と表す．大きさ（ノード数）k のグラフの隣接行列を X_k と表し，第 i ノード第 j ノードとの間のリンクの有無及び種類を表す X_k の要素を x_{ij}，グラフを $G(X_k)$ で表す．ノードの種類を N_p とし，リンクの種類を L_q とする．最小支持度は以下のように定義する．

$$sup(G) = \frac{TrN(G)}{\text{全トランザクションの数}} \qquad (1)$$

ここで，$TrN(G)$ はグラフ G を誘導部分グラフとして含むトランザクションの数である．ここでグラフ G' をグラフ G の部分グラフとし，$V(G), E(G)$ をそれぞれグラフ G のノードの集合，リンクの集合とするとき，任意のノード $u, v \in V(G')$ に対して，$\{u,v\} \in E\{G\} \Leftrightarrow \{u,v\} \in E\{G'\}$ が成り立つとき，G' を G の誘導部分グラフという．最小支持度は従来のバスケット分析と同様に定義し，最小支持度を超える支持度を有するグラフを多頻度グラフと呼ぶ．

対象とするデータはノード及びリンクに種類があり，自己ループを持つ無向・有向グラフである．グラフ構造は隣接行列で表現されるが，どのノードを隣接行列の i 行（i 列）にするかによって隣接行列の表現が変わる．そこで，以下のようにグラフ構造のノードの種類間に例えば辞書順のような順序関係を設け，探索すべき場合の数を抑制する．

$$N_1 < N_2 < \cdots < N_p < \cdots \qquad (2)$$

即ち，あるグラフ構造データのノードで種類の順序が最も若いノードを第1ノードとし，順に第2ノード，第3ノードとする．隣接行列を対称行列に限れば，無向グラフについても同様に扱うことができる．また，以下のように各隣接行列 X_k を Q 進数で表す．ここで Q はリンクの種類数である．

$$code(X_k) = x_{1,2}x_{2,1}x_{1,3}x_{3,1}\cdots x_{k,k-1}x_{k-1,k} \quad (3)$$

ここで各 $x_{i,j}$ は0から Q までの整数である．

従来のApriori同様，隣接行列表現のグラフについても，以下の条件下でのみ2つのグラフを結合しサイズの大きな多頻度グラフの候補を順次生成していく．即ち，グラフの大きさが k の多頻度グラフを2つ考え，その隣接行列を X_k, Y_k とする．

$$X_k = \begin{pmatrix} X_{k-1} & \boldsymbol{x}_1 \\ \boldsymbol{x}_2^T & 0 \end{pmatrix} \quad (4)$$

$$Y_k = \begin{pmatrix} X_{k-1} & \boldsymbol{y}_1 \\ \boldsymbol{y}_2^T & 0 \end{pmatrix} \quad (5)$$

この時以下のように X_k, Y_k を結合し，Z_{k+1} を生成する．

$$Z_{k+1} = \begin{pmatrix} X_{k-1} & \boldsymbol{x}_1 & \boldsymbol{y}_1 \\ \boldsymbol{x}_2^T & 0 & z_{k,k+1} \\ \boldsymbol{y}_2^T & z_{k+1,k} & 0 \end{pmatrix} = \left(\begin{array}{cc|c} \multicolumn{2}{c|}{\multirow{2}{*}{X_k}} & \boldsymbol{y}_1 \\ & & z_{k,k+1} \\ \hline \boldsymbol{y}_2^T & z_{k+1,k} & 0 \end{array}\right) \quad (6)$$

ここで，X_{k-1} は大きさ $k-1$ のグラフの隣接行列，$\boldsymbol{x}_i, \boldsymbol{y}_i$ $(i=1,2)$ は $(k-1)\times 1$ の縦ベクトルである．2つの隣接行列を結合させてできた隣接行列 Z_{k+1} の第1生成行列，第2生成行列の第 k ノードのノードの種類が等しい場合には，上記の X_k を第1生成行列，Y_k を第2生成行列とした場合と，逆に X_k を第2生成行列，Y_k を第1生成行列とした場合では後者の場合が冗長である．そこで，このような冗長な生成を避けるため，以下の関係にある時のみ結合を行う．

$$code(\text{第1生成行列}) \leq code(\text{第2生成行列}) \quad (7)$$

以上のような条件のもとで生成される隣接行列を**正規形**と呼び，非正規形の隣接行列は生成しない．

次に従来のAprioriと類似して，結合してできたグラフ $G(Z_{k+1})$ の第 i ノード $(1 \leq i \leq k-1)$ を開放除去したグラフの隣接行列が全て多頻度グラフを表す隣接行列であれば，それを多頻度グラフの候補と考えられる．ここで，第 i ノードの開放除去とは，第 i ノードとそれにつながるリンクを全て除き，大きさが k の誘導部分グラフを得る操作である．先にも述べたように，このアルゴリズムでは正規形の隣接行列しか探索生成しないために，第 i ノードを開放除去したグラフの隣接行列が正規形でなければ，それが多頻度グラフであるかを過去の探索から容易にチェックする事ができない．よって，詳細は省略するが非正規形の隣接行列の正規化を施しチェックを実施する．また，正規形の中には同じグラフを表す隣接行列が複数存在する場合がある．従って多頻度グラフを得るために各誘導部分グラフの頻度をカウントする際，このような隣接行列に対するカウントを常に合計する必要がある．そこで，同じグラフを表現する正規形の隣接行列のうちコードが最小のものを**正準形**とし，詳細は省略するが正準形とそれへの変換行列を求める．全ての多頻度グラフの候補を取り出し正準形を求めた後，実際にデータベースにアクセスする事によって頻度を計算する．

4·2 数値データの扱い

多次元の数値属性アイテムに関しても理解容易な離散化を与えるために，あくまで属性空間上の各属性軸に垂直な離散化，即ち通常の各属性軸上のしきい値による離散化を行うことを考える．またこの制限内で，極力属性アイテム間の共起情報を失わないよう最低限の粒度の離散化を行うために，Kullback情報量を最小にする赤池情報量基準(AIC)を用いることとした．更に数値属性アイテム間の依存性を考慮すべく，各属性軸個別ではなく全属性空間に亘ってAICを計算し，離散化を行うこととした．AICを最小にする閾値をGreedyに探索し，AICが極小値を過ぎて増加すると離散化を終える[塚田98, 塚田00]．

§1 クラス情報を用いる場合

AICの評価式は以下の通りである．

$$AIC(S_i; i=1,...,r) = 2\sum_{i=1}^{r} |S_i| Ent(S_i) + 2s \quad (8)$$

ただし，$Ent(S_i) = -\sum_{j=1}^{k_i} p_{ij} \log p_{ij}$，$k_i$ はデータ部分集合 S_i に含まれるデータのクラス数，p_{ij} は，部分集合 S_i に含まれるデータの各クラス数の割合，s は数値属性を離散化するしきい値の数，そして，r は属性空間の分割によって得られるデータ部分集合 S_i の総数である．

§2 クラス情報を用いない場合

クラス情報の代わりに，各事例データの各属性値を平均値とする多次元正規確率分布を用いて，情報エントロピーを計算する．属性 x_i についての分散を σ_i^2 とすると，属性空間座標 $X=[x_1, x_2, ..., x_m]$ に事例 $X_k = [x_{1k}, x_{2k}, ..., x_{mk}]$ が存在する確率密度 p_k は，

$$p_k = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp -\frac{(x_i - x_{ik})^2}{2\sigma_i^2} \quad (9)$$

と算出できる．本来なら確率密度 p_k を求める際に複数属性間の依存性を考えるべきであるが，厳密な式をたてることが解析的に困難であること，そして計算量の増加

を防ぐことを考慮して，全属性における同時確率密度 p_k を各属性の確率密度の積で代用する．次に，属性空間の分割によって得られるデータ部分集合 S_i に事例 X_k が存在するエントロピー

$$ENT_{ki} = \int_{V(S_i)} (-p_k \log p_k) dV \qquad (10)$$

を考える．ここで $V(S_i)$ は S_i を全データから切り出す属性部分空間の体積である．各 S_i を切り出す閾値の数 s，データ部分集合の総数 r，事例データ総数 n から AIC は以下のように計算される．

$$AIC(S_i; i = 1, ..., r) = 2\sum_{k=1}^{n}\sum_{i=1}^{r} ENT_{ki} + 2f \qquad (11)$$

第 1 項が近似を含むため，それを考慮して離散化自由度に相当する第 2 項の f は経験的に $f = (1/4)r$ と決められる関数である．

5. 髄膜炎データベースからの知識発見

5·1 クラス情報を用いる場合

髄膜炎診断の共通医療データについて，(case1) 診断を表す記号属性 DIAG をクラスとし，欠損値を含む数値属性 CSF_CELL3 と実際の治療法を表す記号属性 THERAPY2 を除いたデータを用いた場合，(case2) 診断を表す記号属性 Diag2 をクラスとし，欠損値を含む数値属性 CSF_CELL3 と実際の治療法を表す記号属性 THERAPY2 を除いたデータを用いた場合の 2 ケースについて，通常のアイテム集合からなるトランザクションに関する相関ルールマイニング（バスケット分析）を試みた．DIAG は 6 つのクラス (ABCESS, BACTERIA, BACTERIA(E), VIRUS, VIRUS(E), BT(E)) を持つ属性であり，Diag2 は DIAG における ABCESS, BACTERIA, BACTERIA(E) と BT(E) を VACTERIA に，VIRUS と VIRUS(E) を VIRUS にまとめて分類し直した属性である．case1，case2 それぞれについて，クラス付き AIC によって離散化した．

case1 を離散化した時に選ばれる閾値とその AIC 評価値を表 1 に示す．更にバスケット分析を行った時の多頻度アイテム集合 (FI) と相関ルールの数を表 2 に示す．表中の *s-conf*[%] は指定確信度 [%] を表し，*l-sup*[%] は最小指示度 [%] を表す．ルール 1 は得られた相関ルール数であり，更に頭部にクラス (DIAG) を含む相関ルールの数をルール 2 の列に示す．一方，case2 を AIC で離散化したときに選ばれる閾値と AIC 値を表 3 に示す．また，バスケット分析を行ったときの多頻度アイテム集合と相関ルールの数を表 4 に示す．

case1 と case2 について得られた相関ルールを専門家に提示し，以下の四つの指摘を受けた．

(1) AIC による離散区間は適当である．

表 1 case1 の離散化結果

m	属性値	閾値	AIC
1	Cell_Poly	221	260.715
2	Cell_Mono	12	216.906
3	LOC_DATA		176.071
4	EGG_FORCUS		146.235
5	FOCAL		104.699
6	CRP	3.1	72.487
7	Cell_Mono	320	48.143
8	SEX		30.542
9	HEADACHE	3	24.591
10	CSF_GLU	55	22.772
11	BT	37.0	22.000

表 2 case1 を離散化したときのルール数

s-conf[%]	*l-sup*[%]	FI	ルール 1	ルール 2
90	20	895	98	23
	15	1720	176	42
	10	3542	319	76
	2	26755	1732	404
70	20	895	194	63
	15	1720	352	101
	10	3542	610	184
	2	26755	3588	948

表 3 case2 の離散化結果

m	属性値	閾値	AIC
1	Cell_Poly	221	63.781
2	Cell_Mono	15	28.123
3	CT_FIND		16.751
4	CRP	4.2	8.000

表 4 case2 を離散化したときのルール数

s-conf[%]	*l-sup*[%]	FI	ルール 1	ルール 2
90	20	41	10	5
	15	47	10	5
	10	76	11	5
	2	144	19	8
70	20	41	12	6
	15	47	12	6
	10	76	13	7
	2	144	21	10

(2) 得られたルールは医学的に妥当なルールが多い．

(3) 同時に専門家にとって意外なルールも多い．

以下に各ケースについて，専門家にとって興味深いものがあると指摘されたルールの例を示す．なお，各ルールを次のように表示する．

$$\{[\text{属性名}] : [\text{その属性における値の範囲}]\}(\text{条件部}) \Rightarrow \{[\text{属性名}] : [\text{その属性における値の範囲}]\}(\text{結論部}) \qquad (12)$$

$$(sup(\text{支持度})[\%], conf(\text{確信度})[\%])$$

case1

$$\{[\text{HEADACHE}] : [3, 63]], [\text{Cell\_Poly}] : [0, 220], [\text{EEG\_FOCUS}] : -, [\text{LOC\_DAT}] : -\} \Rightarrow \{[\text{DIAG}] : \text{VIRUS}\} \qquad (13)$$

$$(sup = 33.6[\%], conf = 90.4[\%])$$

case2

$$\{[\text{Cell\_Poly}] : [0, 220]\} \Rightarrow \{[\text{CRP}] : [0.0, 4.0]\} \quad (14)$$
$$(sup = 72.9[\%], conf = 95.3[\%])$$
$$\{[\text{Cell\_Mono}] : [0, 12]\} \Rightarrow \{[\text{Cell\_Poly}] : [0, 220]\} \quad (15)$$
$$(sup = 7.9[\%], conf = 91.7[\%])$$

5·2 クラス情報を用いない場合

髄膜炎診断に関する共通データは，欠損値を含む数値属性CSF_CELL3と実際の治療法を表す記号属性THERAPY2を除き，数値属性の数は21，記号属性の数は13，事例数は140で構成される．今回は，数値属性のみを取り上げ，クラスなしAICによる離散化を行った．表5にその結果を示す．総計5つのしきい値による離散化が得られた．

表5 離散化結果

閾値番号	属性	閾値	AICの最小値
1	Cell_Poly	615.2	-8.14
2	WBC	1959.39	-16.22
3	CSF_CELL	633.5	-23.29
4	CSF_GLU	5.2	-28.30
5	LOC	0.26	-29.12
AIC Increase !!!			
6	SEIZURE	0.06	-21.84

次に5つの離散化属性と，属性THERAPY2とDiag2を除いた17個の記号属性の値をアイテムとするデータについて，バスケット分析を行った．以下に導出した相関ルールの一部を示す．

$$\{[\text{Cell\_Poly}] : [0, 615.2], [\text{LOC\_DAT}] : -,$$
$$[\text{SEX}] : \text{F}, [\text{FOCAL}] : -\}$$
$$\Rightarrow \{[\text{WBC}] : [1959.39, 90009], [\text{EEG\_FOCUS}] : -, \quad (16)$$
$$[\text{LOC}] : [0, 0.26], [\text{C\_COURSE}] : \text{negative},$$
$$[\text{CSF\_GLU}] : [5.2, 520], [\text{DIAG}] : \text{VIRUS}\}$$
$$(sup = 17.1[\%], conf = 72.7[\%])$$
$$\{[\text{Cell\_Poly}] : [615.2, 61520], [\text{LOC\_DAT}] : +,$$
$$[\text{SEX}] : \text{M}, [\text{FOCAL}] : -, [\text{C\_COURSE}] : \text{negative}\}$$
$$\Rightarrow \{[\text{WBC}] : [1959.39, 90009], [\text{EEG\_FOCUS}] : -, \quad (17)$$
$$[\text{LOC}] : [0, 0.26], [\text{C\_COURSE}] : \text{negative},$$
$$[\text{WBC}] : [1959.39, 90009], [\text{CSF\_GLU}] : [5.2, 520],$$
$$[\text{DIAG}] : \text{BACTERIA}\}$$
$$(sup = 2.1[\%], conf = 75.0[\%])$$

これら2つのルールより，Cell_Polyの値が低くかつLOC_DATAが陰性であり患者が女性であるとき，ウィルス性髄膜炎([DIAG]:VIRUS)である可能性が高いこと，またCell_Polyの値が高くかつLOC_DATAが陽性であり患者が男性であるとき，細菌性髄膜炎([DIAG]:BACTERIA)である可能性が高いことを推測することができる．

5·3 クラス情報有無の場合の比較

表1と表5を比較すると，選ばれた属性数は目的量を用いる場合の方が多いことがわかる．この理由として，目的量を用いる場合は数値属性と記号属性を離散化の対象としているが，一方で目的量を用いない場合は数値属性のみを対象として離散化を行ったためであることが考えられる．またいずれの場合もCell_Polyという属性が閾値番号1の属性として選択されている．このことから，髄膜炎とCell_Polyという数値属性の間に強い因果関係があることがわかる．

次に相関ルール(13)と(16)を比較すると，条件部に，Cell_Polyの値が低いこと，かつLOC_DATが陰性であることが共通して含まれ，しかもいずれもウィルス性髄膜炎([DIAG]:VIRUS)を結論部として含んでいる．今回用いた髄膜炎に関する医療データについては，目的量を用いる場合と用いない場合では，ほぼ一致した知識が得られる．すなわち，目的量を用いずともある程度専門家にとって興味深いルールを導出することができたと思われる．

6. 膠原病データベースからの知識発見

膠原病に関する共通データは，内科膠原病外来に数年以上通院し，診断・治療・経過の観察が行われている患者のデータを集めたものである．このデータベースは3つのデータ(A,B,C)から構成されており，これらのデータにおいて患者を特定するID番号が，全事例データの先頭に示してある．3つのデータにおいてID番号が同じ事例データであれば同じ患者の症状を表すことから，これらのデータは互いに参照することができる．

Aは，膠原病外来で経過観察されている患者についての基本情報が含まれている．各事例データが一人の患者の症状を表しID番号順に示してある表形式をなすデータである．またBはAと同じ表形式のデータであり，血栓症に関わる特殊検査の情報が含まれている．血栓症は膠原病の合併症として最近注目されている疾患である．更にCは各事例データが測定日とIDの情報を有し，各ID番号が表す一患者に対して測定日によって順序付けられている複数の事例データを持ち，更にそれら複数の事例データがID番号順に並んで構成されている．

6·1 表形式をなす膠原病データの解析

はじめにAから各膠原病を正確に診断するためのパターンの抽出，Bから血栓症の診断に有効な属性の抽出を目的とした解析を行った．これらは表形式データであるので，各属性値をアイテムとしてバスケット分析により相関ルールを導出した．それぞれ離散化の対象として

数値属性のみを取り上げ，数値属性に欠損値を含む事例データを全事例集合から取り除き，クラスなしAICによる離散化を行った．それぞれの離散化結果を表6，表7に示す．何れの場合も3種類の数値属性がそれぞれの閾値によって離散化されることがわかる．表6よりAに関して初診日を表す属性First Dateと誕生日を表す属性Birthdayの閾値が選ばれたことは，膠原病と年齢に相関があると推測される．またBに関しては，表7より選ばれた閾値の全ての属性が抗Cardiolipin抗体の種類を表し，血栓症が抗Cardiolipin抗体と深く関わっていることが推測される．

表6 Aの離散化結果

閾値番号	数値属性	閾値	AICの最小値
0	初期状態		0.0
1	First Date	1994/5/18	-5891.7
2	Description	1995/6/3	-7056.7
3	Birthday	1946/9/14	-11404.0

表7 Bの離散化結果

閾値番号	数値属性	閾値	AICの最小値
0	初期状態		0.0
1	aCL IgA	53.9	-10337.6
2	aCL IgM	207.9	-20675.3
3	aCL IgG	3.3	-28845.1

次に全事例データに関して，離散化された数値属性と全ての記号属性を合わせてバスケット分析を行った．数値離散化の段階では数値属性に欠損値を含む事例データを除いたが，バスケット分析の段階では前段で得られたしきい値を基に欠損値を含む全事例データの数値属性を離散化し解析を行った．多頻度アイテム集合(FI)と相関ルールの数を表8に示す．指定確信度，最小支持度が大きくなる毎に多頻度アイテム集合とルールの数は減少するが，全体的にそれらの数は少ないことがわかる．これは欠損値が多いことがその理由であると考えられる．欠損値が多いと欠損値以外のアイテムを含む集合の支持度が比較的小さくなり，多頻度アイテム集合として選ばれにくい．

表8 多頻度アイテム集合と相関ルールの数

		A		B	
l-sup	s-conf	FI	ルール	FI	ルール
2	50	448	38	654	33
	70		28		23
	90		9		18
5	50	205	23	238	7
	70		11		7
	90		4		4

最後に得られたルールの一部を次に示す．Aの解析から，

$$\{[\text{Birthday}] : [1912/8/28, 1946/9/14],$$
$$[\text{Description}] : [1995/6/3, 1998/12/3],$$
$$[\text{Admission}] : -, [\text{SEX}] : \text{M}\}$$
$$\Rightarrow \{[\text{Diagnosis}] : \text{RA}\} \quad (18)$$
$$(sup = 0.8[\%], conf = 71.4[\%])$$

という相関ルールを得た．ここでBirthdayは生年月日，Descriptionはデータが入力された日，Admissionは入院したかどうか，SEXは性別，Diagnosisは最終診断結果を表す．この相関ルールから比較的年齢が高いことや性別が男性であることが，最終診断が慢性関節リウマチ([Diagnosis]:RA)であることを導くという知識が得られた．また，Bの解析から，

$$\{[\text{aCL\_IgG}] : [3.3, 2150.3], [\text{Diagnosis}] : \text{APS}\}$$
$$\Rightarrow \{[\text{aCL\_IgA}] : [0.0, 53.9], [\text{Thrombosis}] : 1\} \quad (19)$$
$$(sup = 0.6[\%], conf = 55.6[\%])$$

という相関ルールを得た．aCL_IgGは抗Cardiolipin抗体(IgG)，aCL_IgAは抗Cardiolipin抗体(IgA)，Diagnosisは診断結果，Thrombosisは血栓症の有無を表し，この相関ルールから比較的抗Cardiolipin抗体(IgG)が高いことと診断がAPSであることが，比較的抗Cardiolipin抗体(IgA)が低いことと血栓症がもっとも重症である(1はもっとも重症であることを示す)ことを導くという知識が得られた．

6·2 時系列を含む膠原病データの解析

ここでは各膠原病を特徴づける膠原病患者治療歴の時間的パターンの抽出を目的とする解析を行った．そのためにID番号を用いてAの属性Diagnosisを参照し，Cの解析を行った結果を示す．Cは各事例データが治療日とIDの情報を有し，IDと治療日でソートされている．Cを離散化した後，治療日による時間順序と基本情報を含むAにおける診断結果を関係づけたグラフ構造に変換し，AGMにより相関ルールの抽出を行った．

はじめにCについてクラスなしAICによる数値属性の離散化を行った．このデータには欠損値を非常に多く含む属性が多いことから，欠損値が存在する割合が50%以上である数値属性を全事例データから取り除いた．更に残された数値属性に欠損値を含む事例データを全事例データから除き，それらの事例データを対象として離散化を行った．離散化結果を表9に示す．T-BILは総ビリルビン酸濃度，GPTはグルタミン酸ピルビン酸濃度，GOTはグルタミン酸オキサロ酢酸濃度を表す．選ばれた3つの属性のうち2つがグルタミン酸系トランスアミナーゼ濃度を表すことから，これらの濃度が各膠原病と深く関わっていると考えられる．

表9 Cの離散化結果

閾値番号	数値属性	閾値	AICの最小値
0	初期状態		0.0
1	T-BIL	0.87	-55191.3
2	GPT	64.5	-110284.2
3	GOT	49.0	-165220.9

離散化の段階では数値属性に欠損値を含む事例データを全事例データから除いたが，バスケット分析の段階で

は全事例データを対象として解析を行う．ただし，離散化の際に選ばれた属性 (GOT,GPT,T-BIL) の属性値が全て欠損している事例データは，解析の対象から除いた．図1(a) に示されるように，A の最終診断 (Diagnosis) は ID 番号を通じて C の治療歴に対応している．そこで図1(b) のように，A における属性 Diagnosis と属性値の組，そして C における各属性とその属性値の組をノードとし，それに加えて同一治療日であるノードを束ねるためのダミーノードを設ける．次に各ダミーノードとその後日の全ダミーノードの間に時間順序を表すためのリンクを張る．また今回のデータでは治療歴中の治療日 (治療のステップ) が，多い場合には 400 日となる．従ってノードの数が非常に多くなることがあり，計算時間が膨大になる恐れがある．これを防ぐために，1つの治療日を基準として，ある一定のステップ数までを上で述べたようにトランザクションとして構成する．更に患者の治療歴の全ての治療日を基準として上記のトランザクションを得る．こうすることによってトランザクションの数は非常に増加するが，ノードの数は少なく抑えられ計算時間は大きく増大しない．このようなデータ前処理により，トランザクショングラフ数 47652 個，ノードの種類数 216 個のグラフ構造データを得た．

(a) 離散化された，時系列を含むデータ

A

患者番号(ID)	…	診断(Diagnosis)
1	…	RA
2	…	PSS
…	…	…

C

患者番号(ID)	治療日(Date)	GOT	GPT	T-BIL
1	1日目	多い	多い	多い
1	2日目	多い	少ない	多い
…	…	…	…	…
2	1日目	少ない	多い	多い
…	…	…	…	…

(b) 基本的に1人の患者に対して1つのグラフをつくる

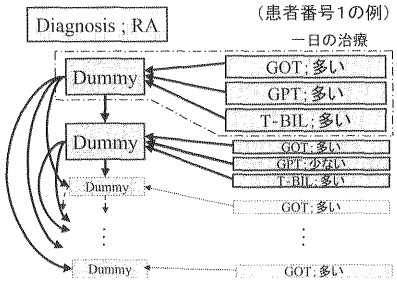

図 1 時系列を含むデータ

以上のようなグラフ構造への変換を行った後，AGM を適用した結果，Pentium300MHz，メモリ 128MB 程度の通常のパーソナルコンピュータ上で，多頻度レベル (support) しきい値を5%に設定し約 20000 秒で全パターンを得ることができた．例として，

- GPT が [0.0,64.5] の範囲を示し，かつ GOT が [3.0,49.0] の範囲を示す事例が2ステップ続くと，診断結果が SLE である (support=5%).
- GPT;[0.0,64.5] かつ GOT;[3.0,49.0] を示す事例が2ステップ続き，更に3ステップ目に T-BIL;[0.1,0.87] を示す事例が存在する (support=59%).
- GPT;[0.0,64.5] かつ GOT；[3.0,49.0] を示す事例の次のステップに，GPT;[0.0,64.5],GOT；[3.0,49.0] そして T-BIL；[0.87,26.1] を示す事例が存在する (support=6.3%).

というようなグラフ構造上の相関ルールが得られた．ただし，専門家からの評価は，取り扱っている時系列長が短すぎるため，医学的知見として有用なパターンが少ないというものであった．

7. 今後の展望

本稿では，命題論理規則で表される程度の複雑さを有する構造を，一般グラフの部分パターン相関ルールという形式で，完全マイニングする AGM 手法について紹介した．本稿では割愛したが，ノード間のリンクに関するデータ前処理によって，ノードが変数ラベルを有するようなパターンのマイニングも可能であることも分かっている．このように一般グラフは命題論理規則を超えて非常に汎用な構造や規則の知識表現であると共に，適切なデータ前処理の実施によって多くの実問題のマイニングを可能とする枠組みを提供する．

AGM は以上のように一般性の高い構造データマイニング手法であるにもかかわらず，従来の帰納論理プログラミングをベースとするマイニング手法に比べれば遙かに高速な処理を実行可能である．前節の最後に示したような時系列パターンのマイニングは，効率的幅優先探索を使用する Dehaspe 等の帰納論理プログラミング手法でさえも非現実的な処理時間を必要とする [Dehaspe 98]．これに対し，AGM では通常のパーソナルコンピュータで約6時間程度を要するのみである．しかしながら，現状の AGM の速度は多くの実規模問題においてまだまだ不十分である．この事例でも専門家にとって意味のあるより長い時系列パターンをマイニングしようとすると，実用的時間内で解を出すことができない．探索の完全性を維持しつつ一層の高速化を図るためには，グラフの数学的性質を利用した効率的枝刈りや対象データに関する領域知識を柔軟に導入できる枠組みが必要とされる．あるいは実用的な範囲内で探索の完全性を諦める枠組みの検討も有効であると考えられる．

一方，本稿では相関ルールマイニングにおける数値データの扱いについても述べた．前処理を中心とする現行の手法は，クラス情報を用いて離散化を行うものと用いないで行うものの2種類がある．何れの手法も相関ルールマイニングの特性を考慮して，離散化結果の理解容易性と属性アイテム間の依存性を考慮している．相関ルール

マイニング自体はクラス情報を用いる手法ではないが，実問題においてはクラスに相当するアイテムが存在する場合も多く，両手法を目的や条件に応じて使い分けることが望ましいと考えられる．ただし，現状のクラス情報を用いない離散化手法においては，AICの第2項fが経験式に留まり，必ずしも常に良好な結果を得るとは限らないことも分かっている．今後，一層の理論面及び実際面からの手法の精密化が必要である．

◇参 考 文 献◇

[Agrawal 94] Agrawal, R. and Srikant, R.: Fast Algorithm for Mining Association Rules in Large Databases. *Proc. of the 20th Very Large Data Bases Conference*, pp.487-499, 1994.

[Agrawal 95] Agrawal, R. and Srikant, R.: Mining sequential patterns. *Proc. of the Eleventh International Conference on Data Engineering (ICDE'95)*, pp.3-14, 1995.

[Arimura 98] Arimura, H., Wataki, A., Fujino, R. and Arikawa, S.: A Fast Algorithm for Discovering Optimal String Pattern in Large Text Database. *Proc. of the 9th International Conference on Algorithmic Learning Theory*, pp.247-261, 1998.

[Dehaspe 98] Dehaspe, L., Toivonen, H. and King, R.D.: Finding frequent substructures in chemical compounds. *Proc. of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pp.30-36, 1998.

[Dzeroski 96] Dzeroski, S.: Inductive Logic Programing for Knowledge Discovery in Database. *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, pp.59-82, Menlo Park, CA, AAAI Press., 1996.

[Inokuchi 99] Inokuchi, A., Washio, T. and Motoda, H.: Basket analysis for graph structured data. *Proc. of PAKDD99: Methodologies for Knowledge Discovery and Data Mining*, pp.420-431, 1999.

[猪口 00] 猪口明博等: 多頻度グラフパターンの完全な高速マイニング手法. 人工知能学会誌, Vol.15, No.6 (掲載予定), 2000.

[Kerber 92] Kerber, R.: Chi Merge: Descretization of Numeric Attributes. *Proc. of the Nineth National Conf. on Artificial Intelligence (AAAI92)*, pp.123-128, 1992.

[King 96] King, R., Muggleton, S., Srinivasan, A. and Sternberg, M.: Structure-activity relationships derived by machine learning; The use of atoms and their bond connectives to predict mutagenicity by inductive logic programming. *Proc. of the National Academy of Sciences*, Vol.93, pp.438-442, 1996.

[Klopman 84] Klopman, G.: Artificial intelligence approach to structure activity studies. *J. Amer. Chem. Soc.*, Vol.106, pp.7315-7321, 1984.

[Klopman 92] Klopman, G.: MultiCASE 1. A hierarchical computer automated structure evaluation program. *QSAR*, Vol.11, pp.176-184, 1992.

[Kramer 97] Kramer, S., Pfahringer, B. and Helma, C.: Mining for causes of cancer: Machine learning experiments at various levels of detail. *Proc. of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pp.223-226, 1997.

[Mannila 97a] Mannila, H. and Toivonen, H.: Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, Vol.1, No.3, pp.241-258, 1997.

[Mannila 97b] Mannila, H., Toivonen, H. and Verkamo, A.I.: Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, Vol.1, No.3, pp.259-289, 1997.

[松本 93] 松本等: 実時間網管理への定性的診断知識の適用手法. 信学技報, AI93-37, pp.9-14, 1993.

[Matsuzawa 00] Matsuzawa, H. and Fukuda, T.: Mining Structured Association Patterns from Database. *Proc. of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp.233-244, 2000.

[Miura 98] Miura, T. and Ishida, T.: Stochastic Node Caching for Memory-bounded Search. *Proc. of Fifteenth National Conference on Artificial Intelligence*, pp.450-456, 1998.

[Quinlan 86] Quinlan, J.R.: Induction of Decision Trees. *Machine Learning*, Vol.1, pp.81-106, 1986.

[Sintani 98] Sintani, T. and Kituregawa, M.: Mining algorithms for sequential patterns in parallel: Hash based approach. *Proc. of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pp.283-294, 1998.

[Srikant 97] Srikant, R., Vu, Q. and Agrawal, R.: Mining Association Rules with Item Constraints. *Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining*, pp.67-73, 1997.

[塚田 98] 塚田誠等: 数値属性離散化における MDLP と AIC の比較. 人工知能学会研究会資料, SIG-KBS-9802-8(1/28), pp.45-52, 1998.

[塚田 00] 塚田誠等: バスケット分析のための AIC による数値属性離散化手法の評価. 人工知能学会研究会資料, SIG-FAI-9904-10(3/22), pp.57-64, 2000.

[Yoda 97] Yoda, K. et al.: Computing Optimized Rectilinear Regions for AssociationRules. *Proc. of Third Int. Conf. on Knowledge Discovery and Data Mining (KDD97)*, pp.96-103, 1997).

[Yoshida 95] Yoshida, K. and Motoda, H.: Clip: Concept Learning from Inference Pattern. *Artificial Intelligence*, Vol.75, No.1, pp.63-92, 1995.

[Wang 97] Wang, K. and Liu, H.: Schema discovery for semistructured data. *Proc. of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pp.271-274, 1997.

2000年7月21日 受理

著 者 紹 介

鷲尾 隆(正会員)は，前掲(Vol.15, No.1, p.186)参照.

元田 浩(正会員)は，前掲(Vol.15, No.1, p.186)参照.