構造データからのアクティブマイニング

研究代表者　元田　浩　　(大阪大学産業科学研究所)
研究分担者　Tu Bao Ho　(北陸先端科学技術大学院大学知識科学研究科)
研究分担者　鷲尾　隆　　(大阪大学産業科学研究所)
研究分担者　矢田 勝俊　　(関西大学商学部)
研究分担者　大原　剛三　(大阪大学産業科学研究所)
研究分担者　吉田　哲也　(北海道大学大学院情報科学研究科)

あらまし　本研究課題の目標は，従来のマイニング手法では扱うことの困難であった構造データ（例えば，化学薬品の構造，蛋白質の構造，Web のリンク構造，多数の検査値からなる時系列医療データ，ゲノムデータ，Web 文書など）に対し，ユーザの価値観を反映した重要なあるいは興味深い部分構造ならびにその特徴を知識として，ユーザの許容時間内に発掘するために必要な基礎技術を開発することである．共通データを主解析対象に，グラフマイニング手法，時系列抽象化手法を開発し，専門家を評価ループに入れたアクティブマイニングによる知識発見の螺旋モデルを実証した．また，これらの技術をビジネス分野や生体情報分野にも展開した．さらに，これらの手法の普及をはかるため，ユーザ指向型マイニングシステム D2MS，強力な前処理機能を持つオープンソースプラットフォーム MUSASHI を開発し，一般に公開した．

1. はじめに

多くのマイニング手法は，マイニングの対象が属性とその値のペアで表現可能であるとの前提に立ち，関係データベースに代表される表形式データを対象にしている．しかし，実際に扱わなければならないデータには表形式では書きにくい，あるいは書けないものも多数ある．例えば，化学薬品の構造，蛋白質の構造，Web のリンク構造，多数の検査値からなる時系列医療データ，ゲノムデータ，Web 文書など枚挙にいとまがない．計画研究「構造データからのアクティブマイニング」の目標は，このような従来のマイニング手法では扱うことの困難であった構造データを効率よくマイニングできるようにする手法を開発し，共通データである，慢性肝炎データと化合物データを始め，種々の構造データに適用し，専門家の助言や評価を受けながら，アクティブマイニングによる知識発見の螺旋モデルを実証することである．この問題に対し，１：構造データをグラフで表現し，一般的なグラフを対象としたマイニング手法を開発する，２：構造データを抽象化して表現する方法を開発し，既存のマイニング手法の適用を可能にする，３．これらの新規手法を内蔵するアクティブマイニングの環境を開発する，の３点から取り組んだ（8）.

2. グラフ構造データに対するマイニング手法

2.1 Graph-Based Induction (GBI) 手法

B-GBI，DT-GBI と慢性肝炎データへの適用　共通データの１つである慢性肝炎データは約 800 個の検査データの 20 年にも及ぶ時系列データである．しかも患者毎に来院の日も検査項目も違い，同一患者でも，等間隔で来院するわけではない．入退院を繰り返す患者もいる．このような多項目の時系列データはそれぞれが独立に変化しているわけではなく，検査項目間には時間的な因果関係がある．それを陽に取り出すことができれば有用な診断知識になり得る．そのため，検査データをグラフに表現し，グラフマイニングの手法により診断知識を取り出すことを試みた．

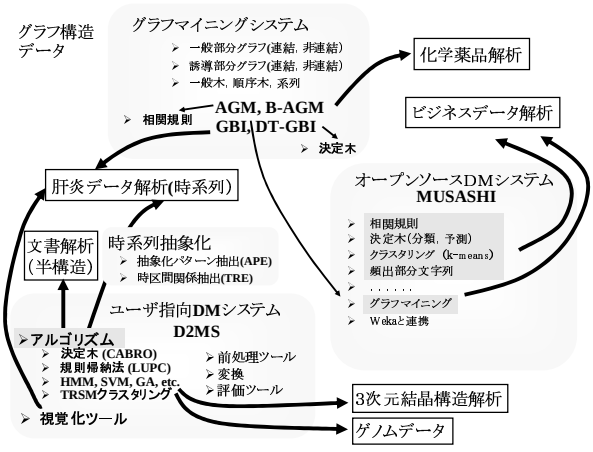


図１　構造データ似関するアクティブマイニングの取り組み

まず，患者毎に各検査値を離散化し，１ヶ月毎に平均し，各検査月を表現するノードからリンクラベルを検査項目，ノードラベルを検査値とする多数の有向グラフを作り，次に各検査月ノードから指定した期間以内の将来の検査月ノードにリンクラベルを検査月間の期間とするリンクをはり，全体として１つの

有向グラフを患者毎に作る．

このようにして作成した多数のグラフから，部分グラフを属性とする決定木を構築する．部分グラフの抽出にはノードペアの逐次チャンキングを基本アルゴリズムとする B-GBI(Beamwise Graph-Based Induction）手法 [11] を利用し，それを属性構築法として再帰的に呼び出す，グラフ構造データ用の決定木作成プログラム DT-GBI を開発した [1] [2].

1 に示す４つの実験を実施し，それぞれに対し，決定木を構築した．実験１，２は肝臓の繊維化の指標（F_0 が正常，F_4 が肝硬変）の予測，実験３は肝炎タイプ B，C の予測，実験４はインターフェロンの効果の予測を目的とするものである．いずれもデータとしては，血液検査，尿検査のみを用いた．データに含まれるノイズを考慮すると，血液検査，尿検査のみからの予測精度としては満足すべきとの見解を専門医から得た．

表 1　DT-GBI による肝炎データ解析結果（予測誤差）

	実験１ F_4 vs. $\{F_0, F_1\}$	実験２ F_4 vs. $\{F_2, F_3\}$
誤差 (%)	12.50	23.52
標準偏差	2.12	2.39
	実験３ B 型 vs. C 型	実験４ IFN R vs. N
誤差 (%)	20.31	22.60
標準偏差	1.57	1.90

2 は実験２に対する決定木の一例である．比較的簡単な判定で肝硬変を予測できることが分かる． 3 と 4 は上２つのノードでテストに使われる検査値のパターンであり，複数の検査値の時間的な相関が捉えられている．これらのパターンの中には専門医にとっても理解困難なものがあり，結果をそのまま受け入れてもらえる段階には至っていないが，アクティブマイニングの有効性は十分認識してもらえた．

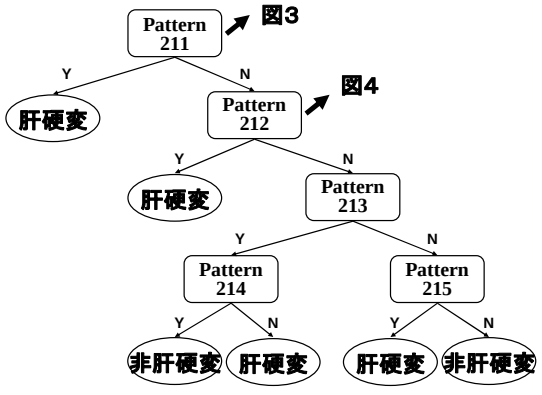


図 2　肝硬変患者同定の決定木の例（実験２）

B-GBI 手法は高速ではあるが重なりのあるパターンを探せない，パターンが存在しても探せないことがあるなどの欠点を持つ．これらの欠点をすべて解消した CI-GBI を最近，新たに開発した [13]．現在，解析を急いでいる．

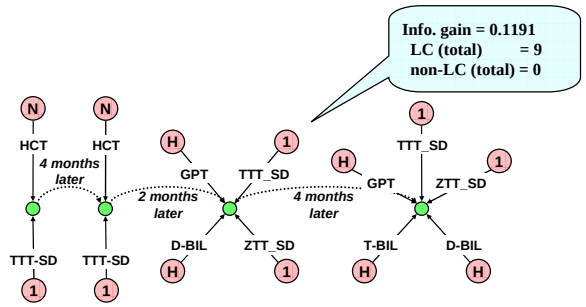


図 3　１段め（ルートノード）の判定パターン

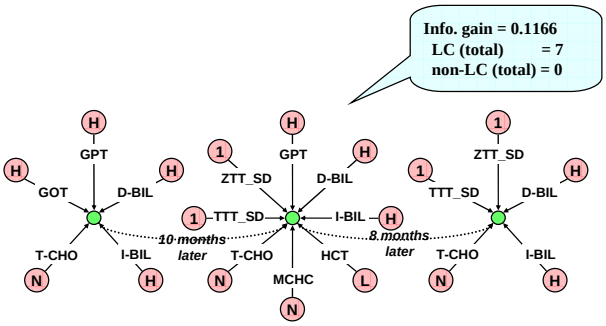


図 4　２段目のノードの判定パターン

2.2　Apriori-based Graph Mining (AGM) 手法

AGM と構造活性相関への適用　構造活性相関解析では，複雑な化学構造式と化学的活性との関係分析が行われる．代表的な手法は QSAR であり，分子構造データをベンゼン環を有するか否か等の構造を特徴づける多数の記述子に変換し，それらと化学的活性の相関解析を行う．カスケードモデル手法は，より直接的に原子鎖に沿う分子構造を属性とし，分類規則を導出する [14]．この手法は化学的活性レベルを特徴づける高精度の相関規則を導出可能である．このような限定した部分構造に着目した手法でも化学構造データは扱えるが，特徴的な部分構造を網羅することはできない．

一般のグラフ構造データに対し多頻度パターン及び相関規則の完全探索が可能な Apriori-based Graph Mining (AGM) 手法を開発した [7]．AGM は指定された頻度閾値を超える多頻度部分グラフを完全探索する．この探索には，多くのグラフ同型問題を解かなければならないため，膨大な計算を要する．5 に示すようにグラフを隣接行列として表現し，部分パターンはそれを含むパターンより出現頻度が高いか少なくとも同じであるという頻度の単調性を利用して問題を解く．グラフ上の第 i ノードを行列の i 行，i 列に対応させ，サイズ（ノード数）k のグラフを $k \times k$ として表す．行列の $\{i, j\}$ 要素で第 i ノードと第 j ノード間のリンクの有無及び種類を表す．第１生成行列で表されるグラフと第２生成行列で表されるグラフの両方が多頻度であり，かつ最下及び最右の行と列のみが異なる行列，すなわち１つのノードのみがトポロジカルな位置が異なる２つグラフをマージし，１サイズ大きな多頻度グラフの候補を生成する．条件を満たすすべての第１及び第２生成行列の合成によって，取りこぼしなく１サイズ大きな多頻度グラフ候補を得ることがで

きる．5 に示すように合成後に元々共通ではなかったノード間にリンクを張るか張らないか，またどの種類のリンクを張るかによって，複数の候補が生成される．候補生成後は，それらとグラフデータを隣接行列の性質を使って効率的に照合し，多頻度か否かを確認する．このようにして小さい多頻度部分グラフからより大きい多頻度部分グラフを，それ以上新たな多頻度部分グラフが見つからなくなるまで逐次探索する．

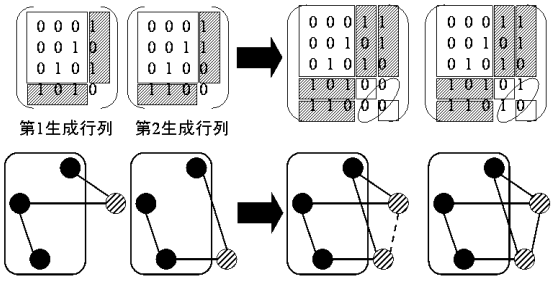


図 5　Apriori-based Graph Mining(AGM) の原理

上述の AGM 手法とカスケードモデル手法を，多数の化合物分子構造の変異原性活性の有無に関するデータに適用した．これはカスケードモデル手法の開発者でかつ化学分野の専門知識が豊富な研究者との共同研究であり，対象データ分野に関する知識が非常に重要な役割を果たした例である [8]．このデータは変異原性活性が無い，低い，中程度，高いという 4 レベルに区分けされている．AGM 手法によってデータ中から多頻度の部分分子構造をマイニングし，それらの部分構造が何れの活性をデータ中に何%もたらすかを調べた．一方，カスケードモデル手法により，原子鎖に沿う特徴によって変異原性活性の 4 レベルに関する分類規則を得た．6 に両者の結果比較の 1 例を示す．

これは多数のマイニング結果パターンから，専門家の分析によって注目すべき組合せを選び，更に考察を加えたものである．左右のパターンの顕著な違いは，負電荷を帯びた NO_2 基側に正電荷を帯びた水素基があるか否かである．水素基がある場合，正電荷に NO_2 基は引かれ，その面はベンゼン環平面に近い配位を取る．従って，この部分の構造が平面状になり遺伝子鎖の二重螺旋内にはまり込みやすくなり，変異原性が高いと考えられる．これに対し後者は，NO_2 基面は必ずしもベンゼン環平面に一致せず，立体障害によって二重螺旋内にはまり込み難いため，変異原性が低いと考えられる．この知見は，専門家にとっても妥当な結果であり，2 次元グラフ表現からこのような知見が得られたことは注目に値する．

消費者行動分析への適用　グラフは複雑なデータ，事象を表現するための強力なデータ構造であり，複雑な因果関係，時系列の変化を伴う社会科学の領域においてもグラフマイニングは重要な示唆を提供できる．顧客の購買履歴データを用いた消費者行動分析では，従来，商品間の購買関連性を明らかにするために相関規則が用いられてきたが，構造を陽に表現できなかった．これに対し，顧客の購買履歴などトランザクションデータをグラフ構造データに変換することによって，情報の消失を極力抑えることができ，既存の手法では得られなかった有用な知

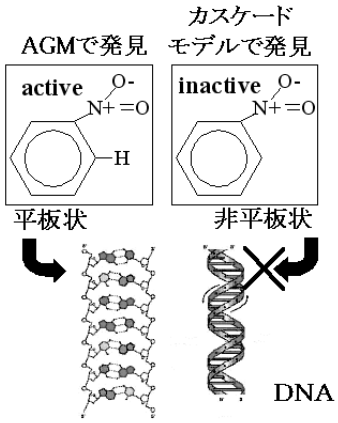


図 6　グラフマイニングによる結果

見を得ることができる [19]．

表 2　購買履歴データの例

顧客 No.	日付	カテゴリ	商品	価格帯
1	04/02-02	ビール	スーパー D	高
1	04/02/02	卵	有機卵	高
1	04/02/02	牛乳	うめぃ牛乳	高
1	04/02/02	マヨネーズ	マヨ Q	普通
1	04/02/12	卵	有機卵	高
1	04/02/12	パン	玄米パン	高
1	04/02/12	牛乳	骨乳	低

2 はある顧客の購買履歴データである．この顧客は 2 月に 2 回来店しており，それぞれ複数のカテゴリの商品を購入している．このような購買行為を 2.1 と同じ手法でグラフ構造データへと変換すると 7 のようになる．1 回の来店で購入される商品群がスター型で表現されており，そのコアノードは購入時点，衛星ノードは購入カテゴリを示しており，購入カテゴリには商品名，価格帯などのラベルがつけられる．複数の購入間はリンクが張られ，購買間隔日数のラベルが付与される．変換後のグラフデータは同時購入されている商品情報だけではなく，来店間隔等の時間情報も含んでおり，このような豊富な情報を持ったデータから特徴的なパターンを抽出できれば，新しい販売促進戦略を計画するための有用な知識発見につながる．実際，日本のスーパーのデータに適用し，有用な知見を発見できた．その一例としてアルコール飲料の購入顧客に関する分析結果を述べる．

日本のアルコール飲料にはビール，発泡酒，チューハイなどのサブカテゴリが含まれている．これらの商品カテゴリの顧客が購入する食料品の傾向をグラフマイニングによって解析し，「ビールのトップシェアを維持している A 社の商品を 10 日以内に 2 回購入する場合に，毎回青果も同時購入しており，かつ鮮魚もしくは精肉も購入する割合が 29 %と非常に高い.」というような規則を得た．A 社の専門家によれば A 社のビールは鮮度を強調した商品であるため，この結果から，購入時，消費者は鮮度に敏感な商品（青果，鮮魚，精肉）を無意識に購入しやすいのではないかという仮説を得ることができた．

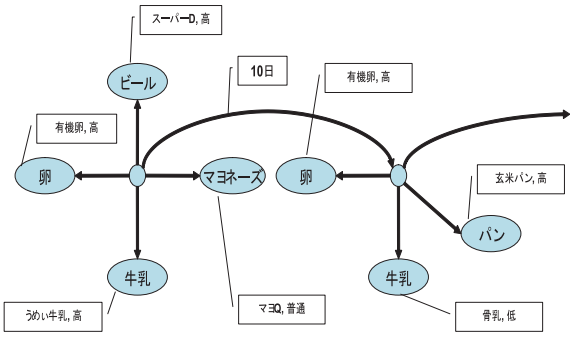


図 7　グラフ構造に変換された購買履歴データ

このような仮説を基礎に新しい販売促進企画が計画された．新しい企画はビールや発泡酒などを冷蔵ショーケースで冷やし，それを青果や鮮魚など様々な売り場に展開することによって，鮮度を訴求した新しい売り場を提案するものであった．この企画を日本のスーパーで実際に実験したところ，ビールや発泡酒などアルコール飲料の売上増加だけではなく，興味深い消費者行動の変化が見られた．この実験中，A 社のビールや発泡酒と同時購入される商品が増え，特に青果や鮮魚が大きく売上を伸ばしたのである．実験店と非実験店との比較，同販促期間中のライバル商品との比較においてもこの傾向が確認され，新しい販売促進企画が新しいビジネスチャンスをもたらしたということができる．

3. 時系列構造データに対するマイニング手法

3.1 時系列抽象化

時系列抽象化（TA）の要点は, 一連のタイムスタンプ付データを抽象化によって時区間を単位とするデータ表現へと変換することにある．不定期な時系列データで構成される肝炎データベースから時区間の概念記述に基く有用なパターンを発見するために 2 段階の手続きからなる時系列抽象化を提案する．第一段階で時系列データに適切な抽象化を施し, 第二段階では抽象化後データに既存データマイニング手法を適用し時区間におけるパターンやモデルを発見する．この第一段階の抽象化に関して, 抽象化パターン抽出 (APE) と時区間関係抽出（TRE）の 2 種類の手法を開発した．

千葉大学病院から提供された元データの検査項目を，専門医が通常重視する検査に絞込んで各患者の履歴データの推移パターンを観察し, 更に専門医と意見交換を行った結果, 時間的変化パターンの特徴が 2 種類に大別できることを見いだした．一方は短期的な変化を見せる検査項目群 (STCT) であり, 炎症などによる肝細胞の破壊を反映し急速に数日から数週間の短期間に劇的に値が変化することを特徴とする．他方は長期的に変化する検査項目群 (LTCT) で, 数ヶ月あるいは数年をかけて緩やかに変化する．主な検査項目の変化パターンの区分を以下に示す．

（1） STCT：GOT, GPT, TTT, ZTT

（2） LTCT：(1) 上昇傾向の検査項目 T-CHO, CHE, ALB, TP, PLT, WBC, HGB, (2) 下降傾向の検査項目 D-BIL, I-BIL, T-BIL, ICG-15

3.2 抽象化パターン抽出法

履歴データの抽象化に際し, 各時系列が STCT, LTCT のいずれかに分類可能なことを踏まえ, 観察に基き典型的変化パターンを決定し，各時系列データから典型パターンへ抽象化するアルゴリズムを開発した．時区間の概念値として典型パターンを表現するために定義した記述構造およびそれを構成する**抽象化基本要素**および要素同士を結合する**関係**は以下の通りである．

抽象化基本要素

（1） *state primitive*: N (正常), L (低値), VL (極低値), XL (超低値), H (高値値), VH (極高値), XH (超高値)

（2） *trend primitive*: S (安定), I(上昇), FI (急速な上昇),

D (下降), FD (急速な下降)

（3） *peak*: P (ピークの有無).

state primitive は, 検査の正常値に対する「状態」を示すもので, 千葉大学病院から提供された閾値を適用して決定した.

関係

（1） > (X>Y): 状態 X から状態 Y へ状態が変化する

（2） & (X&P): 状態 X で且ピークが存在する

（3） – (X–Z): 状態 X の範囲であるが Z の傾向にある

（4） / (X/Y): 主に状態 X であるが状態 Y との推移を頻繁に繰返す

時系列の記述構造は観察を基に以下の４種に集約する.

記述構造パターン

$$
\begin{aligned}
&< pattern > ::= \quad < state\ primitive > \\
&< pattern > ::= \quad < state\ primitive >< relation > | \\
&\qquad\qquad\quad < state\ primitive > | < trend\ primitive > \\
&< pattern > ::= \quad < state\ primitive >< relation >< peak > \\
&< pattern > ::= \quad < state\ primitive >< relation > \\
&\qquad\qquad\quad < state\ primitive >< relation > \\
&\qquad\qquad\quad | < state\ primitive > | < trend\ primitive >
\end{aligned}
$$

以上を用いた抽象化により, 各検査時系列から, 例えば,「ALB = N (ALB の値が常に正常域にある)」,「CHE =H–I" (CHE が高値で継続して上昇傾向にある)」,「GPT =XH&P" (GPT が極めて高値でありピークを伴う)」,「I-BIL =N>L>N" (I-BIL が正常から一旦低値となり, 再び正常に戻った)」, などの概念記述が得られる.

STCT, 特に GPT および GOT においては, 短期間の急激な値の上昇（ピーク）が見られる場合でも, 時系列全体では特定状態に「安定」している事例が多いことから, STCT に対しては, 定常状態を指す**基本状態** *(BS)* と値の急変の有無を示す**ピーク**に基盤を置き特性を記述し [3],LTCT に対しては, 変化自体が長期にわたって緩やかであることから, **状態**と変化の**傾向**の両者に関する情報を含む**状態の変化**を時系列の概念記述の基本とした [9]. 最終的に, 典型パターンを STCT で８種, LTCT で２１種決定し, それぞれに対し，時系列を抽象化する２種類のアルゴリズムを開発した [3].

3.3 時区間関係抽出法

第二の時区間の抽象化手法は, Allen(1983) が提案した時区間論理を基にした時区間関係の抽出（TRE) である. TA が検査ごとの時区間を対象とするのに対し, TRE は検査間の時区間的な関係に着目する.

任意のオブジェクト O_k について, 属性 A_j の時系列 S_{jk} は, 特定時点 t_i の観測値 v_i を用い, $S_{jk} = (v_1, t_1), (v_2, t_2), ..., (v_n, t_n)$ として表現できる. 時区間 $T = (t_s, t_e)$ で発生する事象 E は (E, T) として表す（$t_s, t_e \in \{t_1, t_2, ..., t_n\}$). 時区間論理において Allen は, ２つの事象 A, B 間における時区間の関係を図に示す１３に集約した. こうして得られる時区間関係を結合することで複合的な関係記述が可能となる. 一般に事象 E は各時系列中の関心の対象となる傾向や性質を指すが, 肝炎患者の検査履歴データについては事象を**炎症** もしくは**状態の変化**に限定し, 肝炎問題にとって意味ある時区間関係を重視する. 炎症は高値にある STCT における突発的なピークで特徴付けられることから, E = BS&P の形式でこの事象を取り出す（例「GOT が高値であり且ピークを伴う」). このとき BS は状態要素のどの値であってもよい. LTCT については, 状態の変化が通常は長期的に緩やかに変化することを利用し, LTCT の事象を E = 状態 > 状態 の形式で取り出す（例「ALB が正常から低値へと推移する」). 肝炎データにおける時区間関係抽出手法の概要は以下のとおりである.

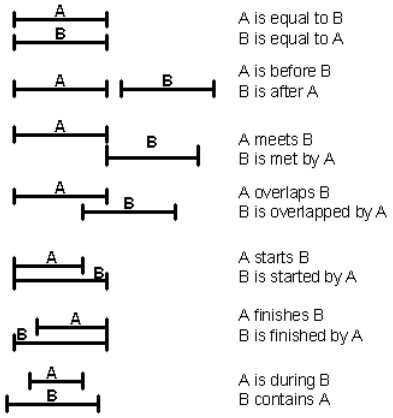


図 8 Temporal relations between two events

（1） 時区間 T において患者 O_k の各検査 A_j の時系列 S_{jk} 中の重要な事象 E (E, T) をすべて見つける.

（2） 患者 O_k の全検査中の全事象について, 時区間論理に基き重要な事象間の時区間関係を検出する.

（3） 抽出した時区間関係のグラフまたはトランザクションを患者 O_k に関する抽象化データとする.

（4） 時区間関係による抽象化データに対しデータマイニング手法を適用して時系列パターンを見つける.

以下は BC 型肝炎データに適用して抽出された時区間関係の相関規則の例である.

Rule 29: "che: H>N"[Overlap]"d-bil: N>H" AND "gpt: VH"[End]"got: H" → HBV (cov: 7, acc: 100%)

Rule 2: "ttt: XH"[Before]"ztt: VH" → HCV (cov: 13, acc = 100%)

Rule 5: "alb: N>L"[Before]"ztt: H" → HCV (cov: 13, acc = 100%)

これ以外の相関規則を総合すると，以下の知見に要約される.

- B 型肝炎の場合, che の正常から高値への上昇に伴い T-CHO あるいは T-BIL が高値から正常へと下降する.
- B 型肝炎の場合, CHE の高値から正常への下降は T-BIL あるいは D-BIL の上昇とほぼ同時である.
- 「ALB が低値から正常へ戻るのに伴い TP が正常値から低値へと下降する」という関係は HBV, HCV 双方で見られるが, このとき同時に「GPT がかなり高い状態が解消した後で GOT が高くなる」のは B 型肝炎の場合である.

- ZTT および TTT は同じタイミングでピークになることが多い尚, C 型肝炎の場合は, ZTT および TTT とも B 型肝炎よりかなり高めの基本状態を示す.
- C 型肝炎の場合, ALB が正常値から低値へと下降した後で様々な他の事象が発生する.

これらに対し医師からは「B 型肝炎に関する規則においてビリルビン (T-BIL, D-BIL, I-BIL) の出現頻度 (6 件中 3 件) は C 型肝炎に関する規則での出現頻度 (10 件中 2 件) よりかなり高い. ビリルビンの動きは T-CHO(総コレステロール), CHE (コリンエステラーゼ) と連動しているようにみえる. これらの変化は肝炎の悪化や回復を意味するが, ビリルビンの肝炎の進行への影響はあまり云われておらず, 上記規則群は B 型肝炎と C 型肝炎の進行の差異を示唆すると考えられる.」とのコメントを得た.

4. 他の構造データに対するマイニング手法

4.1 テキストマイニング

テキストも文法, 単語などの制約下での文字の系列からなる構造データである. 1) **トレランス・ラフ集合モデル** *(TRSM)* に基づくテキスト・クラスタリングおよび情報検索手法, 2) **情報抽出** (IE) 手法および**将来動向予測** (ETD) 手法を提案した. ラフ集合は元々, 等価関係 (reflexive, symmetric, transitive) に基く互いに素なクラスで定義されるが, 種々の関係の拡張が試みられ, トレランス関係 (reflexive, symmetric) もそのひとつである. TRSM はトレランス関係がクラス間の重なりを認めることでテキスト間の意味論的関係を捉えることを可能にする. TRSM に基づき, 階層型と非階層型の 2 種の文書クラスタリング・アルゴリズム およびクラスタを対象とする情報検索 (IR) 手法 [4] を開発した. IR 性能評価用のテキスト集を使った評価および正当性の確認によりこのモデルの長所とりわけ情報検索における精度の向上を確認した.

その他, 情報抽出としてクラスタリングによる同一指示解決問題を取扱う他, 将来動向予測において大規模コーパスから多様な指標の時系列とするトピックの表現モデルを提案した.

4.2 ゲノムデータ解析他

ゲノムデータも代表的な空間構造データである. サポートベクタマシンによるタンパク質の β-ターンおよび γ-ターンの予測・分析 [16], 相関規則マイニング技法による酵母菌の転写因子結合領域データおよび遺伝子発現データからの転写調節モジュールの発見などを試みた [15]. 主要成果としては生物学的課題への適切なマイニング手法の選定に成功したこと, ゲノムデータの表現スキーマを決定したことの他, これらの手法を適用して興味深い知識を発見したことである. その他, 2 次元の化学構造における類似性に関する基礎研究も行い, 経路分析に応用した [10].

5. アクティブマイニングツール

5.1 Data Mining with Model Selection (D2MS)

利用者のマイニングプロセスへの関与を支援するユーザ指向型データマイニングシステムとしてモデル選択を視覚的に支援する D2MS (Data Mining with ModelSelection) を開発した. 3. の時系列抽象化では, 抽象化データのマイニングに D2MS を利用している. 利用者はこのシステムにおいて設定を変えることで多様なアルゴリズムの組合せを試み, その結果の比較評価することが可能である [3], [5], [6]. D2MS の特徴を以下にまとめる.

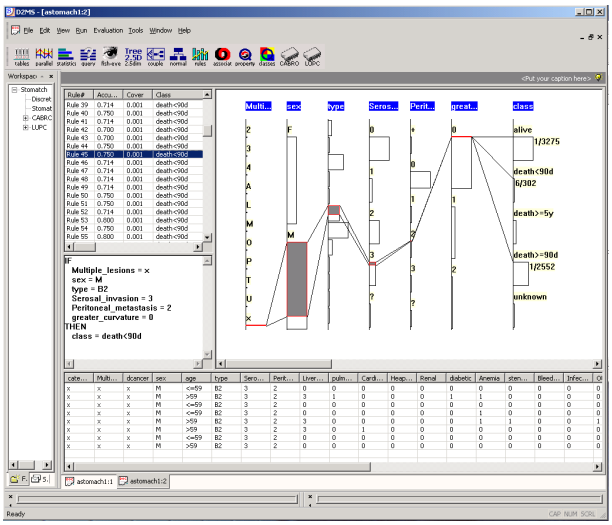


図 9 Rule visualization in D2MS

（1） 性能メトリックスと視覚化機能の提供により, 利用者による定量的・定性的なモデルの評価を支援する.

（2） 視覚化（データ, 規則, 決定木）により利用者の対象領域に関する理解および領域専門家によるマイニング結果の評価を支援する.

（3） 利用者の持つ背景知識の知識発見プロセスへの組込みを支援する. 例えば, 利用者の関心に応じて特定の属性と値の組合せに特化した規則, あるいはそうでない規則の発見を可能とする.

5.2 MUSASHI

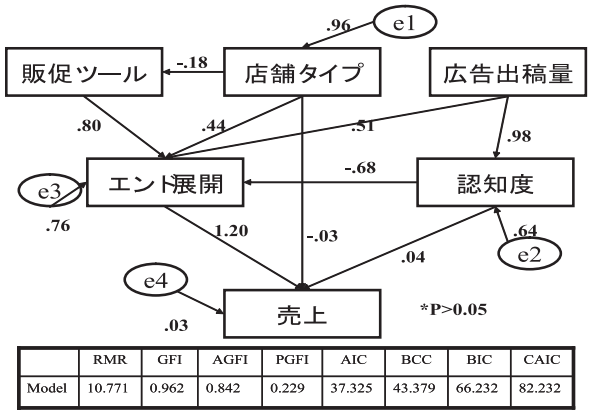
大量データの効率的処理, ユーザーの要求への柔軟な対処を目指したデータマイニングプラットフォーム MUSASHI（Mining Utilities and System Architecture for Scalable processing of HIstorical data）を開発した [12]. MUSASHI はデータマイニングを支えるシステムアーキテクチャであり, 知識発見プロセスで最も労力を必要とする前処理にその強みがある. リレーショナルデータベースやデータウェアハウスの導入を必要とせず, XML で記述された大量データの効率的処理を可能にする. MUSASHI はオープンソースシステムとして開発しており, 誰でも [12] から自由にダウンロードし利用することが可能である.

MUSASHI は, XML で記述されたデータを処理対象とする. XML（eXtensible Markup Language：拡張可能なマークアップ言語）とは, W3C(World Wide Web Consortium) にて策定されたデータ記述言語で, 利用者が自由にタグを定義し, 文書構造を柔軟に表現できる特徴を有す. 例えば MUSASHI が提案する XML テーブルは XML による表形式のデータ構造であり, 一般的なトランザクションデータの処理に効果を発揮する.

MUSASHI は UNIX で伝統的に主張されてきた考え方を踏襲し，大量の XML データに対して単一の機能を持つコマンドを組み合わせ，シェルスクリプトとして実装する．そして，RFM 分析 (Recency, Frequency, Monetary) のように頻繁に利用されるシェルスクリプトはモジュールとして組み込まれ，ユーザーは単一のコマンドと同様に利用することができるようになっている．このような対処法は多様なユーザー要求に対する柔軟性が高く，アプリケーションの開発時間，コストを大幅に削減することができる．MUSASHI は基本的なデータ操作だけではなく，相関規則，決定木，グラフマイニングなどのデータマイニングコマンドも内包しており，オペレーションログの蓄積，大規模データの高速な前処理，規則抽出などをすべて含んだデータマイニングプラットフォームになっている．

現在，MUSASHI のアプリケーションとしては，流通業での利用を前提にした CRM システム (Customer Relationship Management)，C-MUSASHI，製造業，流通業における消費者調査システム，CODIRO が開発されている．C-MUSASHI [18] はデータウェアハウスなどを新たに導入することなく膨大な顧客の購買履歴を蓄積・処理し，効率的な顧客管理を可能にするための MUSASHI 上の CRM システムである．C-MUSASHI は POS レジが出力するログデータを直接，XML データに変換し蓄積することができる．システムには会計情報などの基本的なデータ処理を扱う店舗営業に利用される価格管理なども含まれている．独自の顧客管理ツールとしてはデータマイニング技術を利用した優良顧客のバスケット分析モジュールや離反防止分析モジュールなどが内蔵されている．

CODIRO [17] は企業内で蓄積されているデータベース，WEB 上のテキスト情報，モバイルネットワークの顧客データなど多様なデータベースを統合した消費者調査システムである．テレビ CM の広告効果のように様々な要素が複雑に関係する消費者行動への影響をモデル化することができる．セット米飯商品を対象にした実験結果を行ったところ，テレビ CM 出稿後，エンドと呼ばれる売場の展開方法の最適なタイミングが発見できた．また 10 のような広告，消費者認知，店頭情報を取り入れた販売効果モデルを構築することができた．



	RMR	GFI	AGFI	PGFI	AIC	BCC	BIC	CAIC
Model	10.771	0.962	0.842	0.229	37.325	43.379	66.232	82.232

図 10　広告の販売効果モデル

6. おわりに

以上，構造データに対するアクティブマイニングの取り組みと 4 年間の成果概要を報告した．構造データを表現可能な一般グラフデータからのマイニング手法，時系列データを抽象化して表形式データに変換する手法の開発を中心に，共通データ（慢性肝炎，化学薬品）を対象にアクティブマイニングの螺旋モデルを実証した．さらに得られた手法を内蔵するユーザ指向型のマイニングシステム D2MS，オープンソースプラットフォーム MUSASHI を開発し一般に公開した．加えて，ビジネスデータ，文書データ，ゲノムデータなどの構造データにも目を向け，新しい有望なアプローチを提案するなど，多くの成果を得た．当初の目標は達成出来たものと考える．今後は，開発技術の洗練化と他分野への普及に注力する．グラフマイニングに関しては，その数学的性質上計算量の壁の存在は避けられず，大規模なグラフに対しては，まだ計算時間の点で問題を有す．ヒューリスティックスや積極的な領域知識の導入などの工夫が必要であり，引続き検討する．

文　献

[1] G. Geamsakul, T. Matsuda, T. Yoshida, H. Motoda, and T. Washio. Classifier construction by graph-based induction for graph-structured data. In *Proc. of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (Springer Verlag LNAI2637)*, pp. 52–62, 2003.

[2] G. Geamsakul, G. T. Yoshida, K. Ohara, H. Motoda, H. Yokoi, and K. Takabayashi. Constructing a decision tree for graph-structured data and its applications. *Fundamenta Informaticae*, Vol. **, No. 1, pp. ***–***, 2005. (in press).

[3] T.B. Ho and D.D. Nguyen. Chance discovery and learning minority classes. *Journal of New Generation Computing*, Vol. 21, No. 2, pp. 147–160, 2003.

[4] T.B. Ho and N.B. Nguyen. Nonhierarchical document clustering by a tolerance rough set model. *International Journal of Intelligent Systems*, Vol. 17, No. 2, pp. 199–212, 2002.

[5] T.B. Ho, T.D. Nguyen, and D.D. Nguyen. Visualization support for a user-centered kdd process. In *Proc. of ACM International Conference on Knowledge Discovery and Data Mining KDD-02*, pp. 519–524, 2002.

[6] T.B. Ho, T.D. Nguyen, H. Shimodaira, and M. Kimura. A knowledge discovery system with support for model selection and visualization. *Applied Intelligence*, Vol. 19, , 2003.

[7] A. Inokuchi, T. Washio, and H. Motoda. Complete mining of freqent patterns from graphs: Mining graph data. *Machine Learning*, Vol. 50, pp. 321–354, 2003.

[8] A. Inokuchi, T. Washio, T. Okada, and H. Motoda. Applying the apriori-based graph mining method to mutagenesis data analysis. *Journal of Computer Aided Chemistry*, Vol. 2, pp. 87–92, 2001.

[9] S. Kawasaki, T.D. Nguyen, and T.B. Ho. Temporal abstraction for long-term changed tests in the hepatitis domain. *Journal of Advanced Computational Intelligence & Intelligent Informatics*, Vol. 17, No. 3, pp. 348–354, 2003.

[10] S.Q. Le and T.B. Ho. A novel graph-based similarity measure for 2d chemical structures. *Genome Informatics*, Vol. 14, No. 2, 2004.

[11] T. Matsuda, T. Yoshida, H. Motoda, and T. Washio. Mining patterns from structured data by beam-wise graph-based induction. In *Proc. of The Fifth International Conference on Discovery Science*, pp. 422–429, 2002.

[12] Musashi-project, 2004. http://musashi.sourceforge.jp/.

[13] P.C. Nguyen, K. Ohara, H. Motoda, and T. Washio. Cl-

gbi: A novel strategy to extract typical patterns from graph data. In *Proc. of Joint Workshop of Vietnamese Society of AI, SIGKBS-JSAI, ICS-IPSJ and IEICE-SIGAI on Active Mining*, pp. **–**, 2004. (in press).

[14] T. Okada. Datascape survey using the cascade model. In *Proc. of Discovery Science 2002*, pp. 233–246. Springer, 2002.

[15] T.H. Pham, K. Satou, and T.B. Ho. Mining yeast transcriptional regulatory modules from factor dna-binding sites and gene expression data. *Genome Informatics*, Vol. 14, No. 2, 2004.

[16] T.H. Pham, K. Satou, and T.B. Ho. Prediction and analysis of beta-turn and gama-turns in proteins by support vector machine. *Genome Informatics*, Vol. 14, No. 1, pp. 196–205, 2004.

[17] K. Yada, Y. Hamuro, N. Katoh, and K. Kishiya. The future direction of new computing environment for exabyte data in the business world. In *Proc. of SAINT 2005*, 2005. to appear.

[18] K. Yada, Y. Hamuro, N. Katoh, T. Wachio, I. Fusamoto, D. Fujishima, and T. Ikeda. Data mining oriented crm systems based on musashi: C-musashi. In *Proc. of Second International Workshop on Active Mining*, pp. 52–61, 2003.

[19] K. Yada, H. Motoda, T. Washio, and A. Miyawaki. Consumer behavior analysis by graph mining technique. In *Proc. of KES 2004, LNAI3214*, pp. 800–806. Springer, 2004.