

特集 「アクティブマイニング」

# 構造データからのアクティブマイニング

## Active Mining for Structured Data

元田 浩  
Hiroshi Motoda  
大阪大学産業科学研究所  
The Institute of Scientific and Industrial Research, Osaka University.  
motoda@ar.sanken.osaka-u.ac.jp

ホー ツー バオ  
Tu Bao Ho  
北陸先端科学技術大学院大学知識科学研究科  
School of Knowledge Science, Japan Advanced Institute of Science and Technology.  
bao@jaist.ac.jp

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

矢田 勝俊  
Katsutoshi Yada  
関西大学商学部  
Faculty of Commerce, Kansai University.  
yada@ipc.ku.kansai-u.ac.jp

吉田 哲也  
Tetsuya Yoshida  
北海道大学大学院情報科学研究科  
Graduate School of Information Science and Technology, Hokkaido University.  
yoshida@meme.hokudai.ac.jp

大原 剛三  
Kouzou Ohara  
大阪大学産業科学研究所  
The Institute of Scientific and Industrial Research, Osaka University.  
ohara@ar.sanken.osaka-u.ac.jp

**Keywords:** active mining, structured data, graph mining, temporal abstraction, mining tools.

### 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としている。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、タンパク質の構造、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」はA02班「ユーザ指向アクティブマイニング」に属す四つの計画研究のうちの一つである。本計画研究の目標は、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良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はじめ、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らせん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、

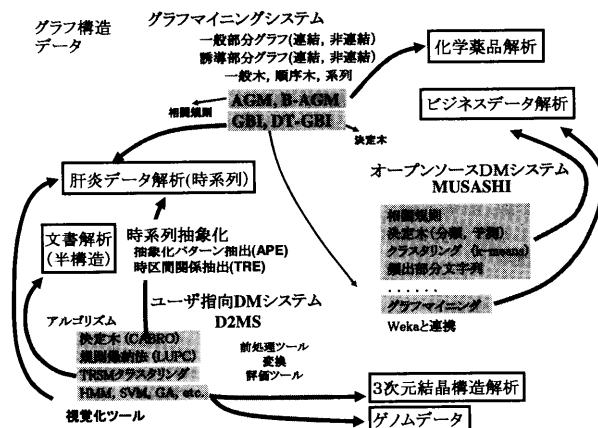


図1 構造データに関するアクティブマイニングの取組み

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(図1)。

### 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 共通データの一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。しかも患者ごと



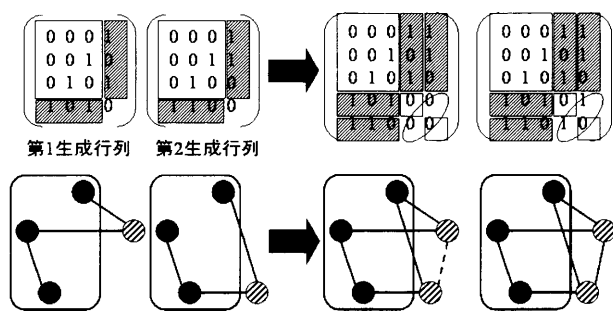


図5 Apriori-based Graph Mining (AGM) の原理

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および種類を表す。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、かつ最下および最右の行と列のみが異なる行列、すなわち一つのノードのみがトポロジカルな位置が異なる二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 および第 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。図 5 に示すように合成後にもともと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。

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 [Inokuchi 01]。このデータは変異原性活性がない, 低い, 中程度, 高いという 4 レベルに区別されている。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。図 6 に両者の結果比較の一例を示す。

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、さら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  $\text{NO}_2$  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  $\text{NO}_2$  基は引かれ、

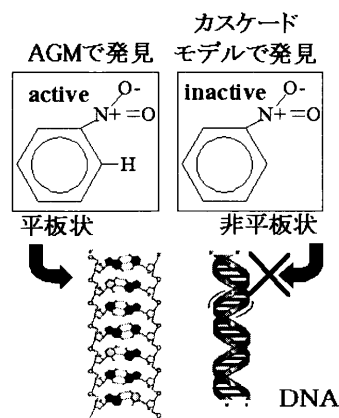


図6 グラフマイニングによる結果

その面はベンゼン環平面に近い配位を取る。したがって、この部分の構造が平面状になり遺伝子鎖の二重らせん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、 $\text{NO}_2$  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らせん内にはまり込みにくいいため、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見を得ることができる [Yada 04]。

表 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ーの商品を購入している。このような購買行為を 2・1 節と同じ手法でグラフ構造データへと変換すると図 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ーを示しており、購入カテゴリーには商品名、価格帯などのラベルが付けられる。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なく、来店間隔などの時間情報も含んでおり、このような豊富な情報をも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

表 2 購買履歴データの例

顧客No.	日付	カテゴリー	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	うめい牛乳	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

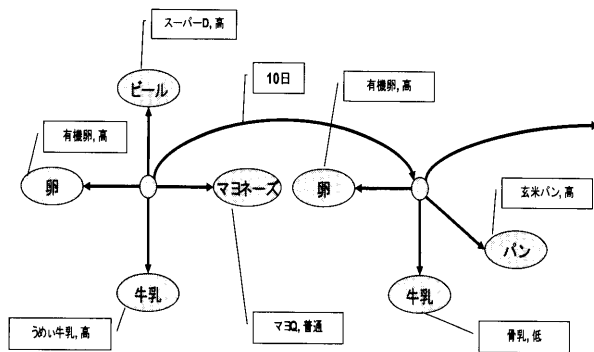


図 7 グラフ構造に変換された購買履歴データ

イなどのサブカテゴリーが含まれている。これらの商品カテゴリー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 回購入する場合には、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この仮説を得ることができた。

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さまざま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に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き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づく有用なパターンを発見するために二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り込んで各患者の履歴データの推移パターンを観察し、さら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、数か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 の何れかに分類可能なことを踏まえ、観察に基づき典型的变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どうしを結合する関係は以下のとおりである。

##### 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しきい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 でかつピークが存在する

(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り返す

時系列の記述構造は観察をもとに以下の 4 種に集約する。

記述構造パターン

<pattern> ::= <state primitive>  
 <pattern> ::= <state primitive><relation> |  
                   <state primitive> | <trend primitive>  
 <pattern> ::= <state primitive><relation><peak>  
 <pattern> ::= <state primitive><relation>  
                   <state primitive><relation>  
                   | <state primitive> | <trend primitive>

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 が正常からいったん低値となり、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 [Ho 03a], 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 [Kawasaki 03]. 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 [Ho 03a].

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もと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、 $S_{jk} = (v_1, t_1), (v_2, t_2), \dots, (v_n, t_n)$  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 ( $t_s, t_e \in \{t_1, t_2, \dots, t_n\}$ )。時区間論理において Allen は、二つの事象 A, B 間における時区間の関係を図 8 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づけられることから、 $E = BS\&P$  の形式でこの事象を取り出す (例「GOT が高値でありかつピークを伴う」)。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  $E = \text{状態} > \text{状態}$  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

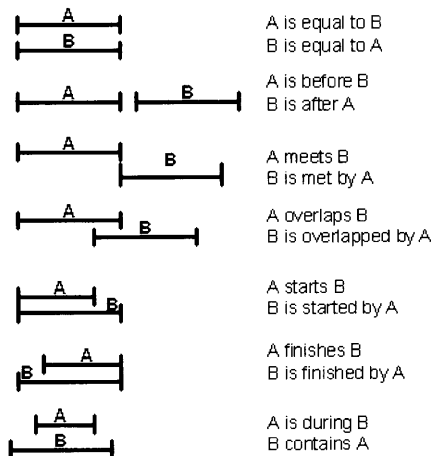


図 8 Temporal relations between two events

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。

- (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  $E (E, T)$  をすべてを見つける。
- (2) 患者  $O_k$  の全検査中の全事象について、時区間論理に基づ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: “ttr: 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。なお、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

## 構造データからのアクティブマイニング

た後でさまざまなほかの事象が発生する。

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見える。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いわれておらず、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思われる」とのコメントを得た。

#### 4. ほか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。ラフ集合はもともと、等価関係 (reflexive, symmetric, transitive) に基づ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、トランス関係 (reflexive, symmetric) もその一つである。TRSMはト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關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[Ho 02a]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り扱うほか、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 ゲノムデータ解析など

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 $\beta$ -ターンおよび $\gamma$ -ターンの予測・分析[Pham 04b], 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[Pham 04a]。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ほか、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[Le 04]。

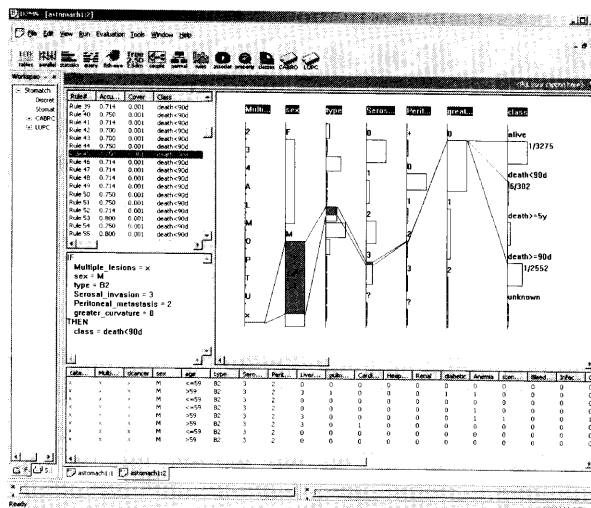


図9 Rule visualization in D2MS

#### 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 Model-Selection) を開発した。3章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[Ho 02b, Ho 03a, Ho 03b]。D2MSの特徴を以下にまとめる。

- (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) 利用者のもつ背景知識の知識発見プロセスへの組み込みを支援する。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[Musashi-project 04]。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[Musashi-project 04]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も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く。このような対処法は多様なユーザ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度前処理、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[Yada 03]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[Yada 05]は企業内で蓄積されているデータベ

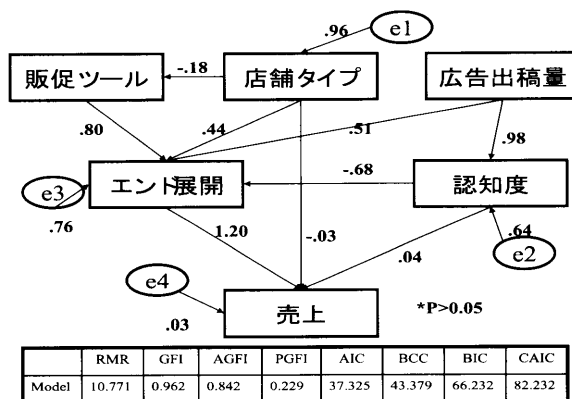


図10 広告の販売効果モデル

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さまざま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。また図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。

## 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ラセンモデルを実証した。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した。今後は、開発技術の洗練化と他分野への普及に注力する。

## 謝辞

本研究の一部は文部省科学研究費特定領域研究(13131206)の補助による。最後に、研究の一部でご協力いただいた北陸先端科学技術大学院大学知識科学研究科河崎さおり助手、肝炎データ解析に際し有益なご助言をいただいた千葉大学医学部高林克日己教授、横井英人助手に感謝します。

## ◇ 参考文献 ◇

- [Geamsakul 03] Geamsakul, G., Matsuda, T., Yoshida, T., Motoda, H. and Washio, T.: Classifier Construction by Graph-Based Induction for Graph-Structured Data, *Proc. 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (Springer Verlag LNAI2637)*, pp. 52-62 (2003)
- [Geamsakul 05] Geamsakul, G., Yoshida, G. T., Ohara, K., Motoda, H., Yokoi, H. and Takabayashi, K.: Constructing a Decision Tree for Graph-Structured Data and its Applications, *Fundamenta Informaticae*, Vol. 66, No. 1-2 (2005) (in press)
- [Ho 02a] Ho, T. and Nguyen, N.: Nonhierarchical Document Clustering by a Tolerance Rough Set Model, *Int. Journal of Intelligent Systems*, Vol. 17, No. 2, pp. 199-212 (2002)
- [Ho 02b] Ho, T., Nguyen, T. and Nguyen, D.: Visualization Support for a User-Centered KDD Process, *Proc. ACM Int. Conference on Knowledge Discovery and Data Mining KDD-02*, pp. 519-524 (2002)
- [Ho 03a] Ho, T. and Nguyen, D.: Chance Discovery and Learning Minority Classes, *Journal of New Generation Computing*, Vol. 21, No. 2, pp. 147-160 (2003)
- [Ho 03b] Ho, T., Nguyen, T., Shimodaira, H. and Kimura, M.: A Knowledge Discovery System with Support for Model

Selection and Visualization, *Applied Intelligence*, Vol. 19 (2003)

- [Inokuchi 01] Inokuchi, A., Washio, T., Okada, T. and Motoda, H.: Applying the Apriori-based Graph Mining Method to Mutagenesis Data Analysis, *Journal of Computer Aided Chemistry*, Vol. 2, pp. 87-92 (2001)
- [Inokuchi 03] Inokuchi, A., Washio, T. and Motoda, H.: Complete Mining of Frequent Patterns from Graphs: Mining Graph Data, *Machine Learning*, Vol. 50, pp. 321-354 (2003)
- [Kawasaki 03] Kawasaki, S., Nguyen, T. and Ho, T.: 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)
- [Le 04] Le, S. and Ho, T.: A novel graph-based similarity measure for 2D chemical structures, *Genome Informatics*, Vol. 14, No. 2 (2004)
- [Matsuda 02] Matsuda, T., Yoshida, T., Motoda, H. and Washio, T.: Mining Patterns from Structured Data by Beam-wise Graph-Based Induction, *Proc. 5th Int. Conference on Discovery Science*, pp. 422-429 (2002)
- [Musashi-project 04] Musashi-project (2004), <http://musashi.sourceforge.jp/>
- [Nguyen 04] Nguyen, P., Ohara, K., Motoda, H. and Washio, T.: Cl-GBI: A Novel Strategy to Extract Typical Patterns from Graph Data, *Proc. Joint Workshop of Vietnamese Society of AI, SIGKBS-JSAI, ICS-IPSJ and IEICE-SIGAI on Active Mining*, pp. 105-110 (2004)
- [Okada 02] Okada, T.: Datascape Survey using the Cascade Model, *Proc. Discovery Science 2002*, pp. 233-246, Springer (2002)
- [Pham 04a] Pham, T., Satou, K. and Ho, T.: Mining yeast transcriptional regulatory modules from factor DNA-binding sites and gene expression data, *Genome Informatics*, Vol. 14, No. 2 (2004)
- [Pham 04b] Pham, T., Satou, K. and Ho, T.: Prediction and Analysis of beta-turn and gamma-turns in Proteins by Support Vector Machine, *Genome Informatics*, Vol. 14, No. 1, pp. 196-205 (2004)
- [Yada 03] Yada, K., Hamuro, Y., Katoh, N., Wachio, T., Fusamoto, I., Fujishima, D. and Ikeda, T.: Data Mining Oriented CRM Systems Based on MUSASHI: C-MUSASHI, *Proc. 2nd Int. Workshop on Active Mining*, pp. 52-61 (2003)
- [Yada 04] Yada, K., Motoda, H., Washio, T. and Miyawaki, A.: Consumer Behavior Analysis by Graph Mining Technique, *Proc. KES 2004, LNAI3214*, pp. 800-806, Springer (2004)
- [Yada 05] Yada, K., Hamuro, Y., Katoh, N. and Kishiya, K.: The Future Direction of New Computing environment for Exabyte Data in the Business World, *Proc. SAINT 2005* (2005), to appear.

2005年1月5日 受理

## 著者紹介

元田 浩 (正会員)



1965年東京大学工学部原子力工学科卒業。1967年同大学院原子力工学専攻修士課程修了。同年、日立製作所に入社、1995年退社。現在、大阪大学産業科学研究所教授(知能システム科学研究部門)。機械学習、知識獲得、知識発見、データマイニングなどの研究に従事。工学博士。1977、1984年日本原子力学会論文賞、1991、1994、2002年人工知能学会論文賞受賞。2001年人工知能学会業績賞受賞、ほか受賞10件、情報処理学会、日本ソフトウェア科学会、日本認知科学会、AAAI、IEEE Computer Society各会員。

ホー ツー バオ (正会員)



1978年Hanoi University of Technology工学部卒業。1987年パリ第6大学博士課程修了。博士。1998年より北陸先端科学技術大学院大学知識科学研究科教授(知識創造論)。機械学習、知識ベースシステムおよび知識発見の基礎研究を中心に、応用としてバイオインフォマティクスおよび材料科学における計算手法に従事。

鷲尾 隆 (正会員) は、前掲 (Vol. 20, No. 1, p. 66) 参照。

矢田 勝俊 (正会員) は、前掲 (Vol. 20, No. 1, p. 66) 参照。

吉田 哲也 (正会員)



1997年東京大学大学院博士課程修了。工学博士。同年、大阪大学大学院基礎工学研究科助手。現在、北海道大学大学院情報科学研究科助教授(コンピュータサイエンス専攻)。機械学習、知識獲得、データマイニングなどの研究に従事。情報処理学会、ヒューマンインタフェース学会各会員。

大原 剛三 (正会員)



1995年大阪大学大学院基礎工学研究科前期課程修了。1997年大阪大学産業科学研究所助手、現在に至る。1996年日本学術振興会特別研究員。博士(工学)。機械学習、データマイニング、および非単調推論に関する研究に従事。IEEE、AAAI、電子情報通信学会、情報処理学会各会員。