人間・マシン協調型データマイニング

磯部　成二

1．はじめに

データウェアハウス等の大量データからビジネスの成功に結びつく有益な知識を得るための技術としてデータマイニングが注目されている．これまで，ビジネス分野のデータ分析には，商品売上データを商品別，販売店別，月別といった多次元の角度から分析するOLAP(Online Analytical Processing)のような人間操作による手法と，顧客を分類しダイレクトメールに反応を予測するクラスタリングやクラシフィケーション，よく併買される商品の組合せパターンを求める相関ルール等のアルゴリズムによる手法が用いられていた．

この分野のデータ分析には，①大量データの高速な処理，②定量的な判断材料の提供，③非定型な分析要求への対応，④人間による解釈・再試行の支援といった多様な要求がある．しかし，人間操作手法では①，②に，アルゴリズム手法では③，④に難があり，現状の手法では，仮説検証や定型的ルールの発見には対応できるが，新しい仮説や新しいルールの発見には十分対応できないという問題があった．

そこで，人間が優れている創造力やパターン認識能力と，コンピュータが優れている記憶・計算能力の長所を活かし，両者の欠点を補完することにより上記の要求を満足する人間・マシン協調型のデータマイニングシステムが実現できると考えられる．このシステムの実現に当たっては，人間とマシンの介在を可能にするインタフェース技術として視覚化技術が重要である．

本稿では，データマイニングの全プロセスにデータ視覚化技術を適用し，プロセスの制御を人間が，ルール候補の生成をマシンが，その結果解釈を人間が行うことを基本とする人間・マシン協調型データマイニングシステムの構成法と，その適用例について述べる．

いそべ　せいじ　NTT　情報通信研究所
〒239-0847　横須賀市光の丘１－１

2．データマイニングプロセスと要求条件

2.1　データマイニングのプロセス

データマイニングにおける知識発見のプロセスは，①アプリケーション領域の学習，②ターゲットデータ集合の生成，③データクリーニングと前処理，④データ削減と射影，⑤データマイニングの目標設定，⑥マイニングアルゴリズムの選択，⑦マイニングアルゴリズムの実行，⑧マイニング結果の解釈と知識の精練，⑨発見された知識の使用の９ステップから構成される[1]．このプロセスは，人間が中心となってマシン処理を選択的に呼び出し，目的の結果を導出する対話的で反復的な作業であり，人間とマシンの分担，および人間とマシンのコミュニケーションが重要な課題である．

2.2　プロセスと要求条件

上記プロセスは，図１に示すように事前準備，マイニング，知識精練の３つのステップに分類される．以

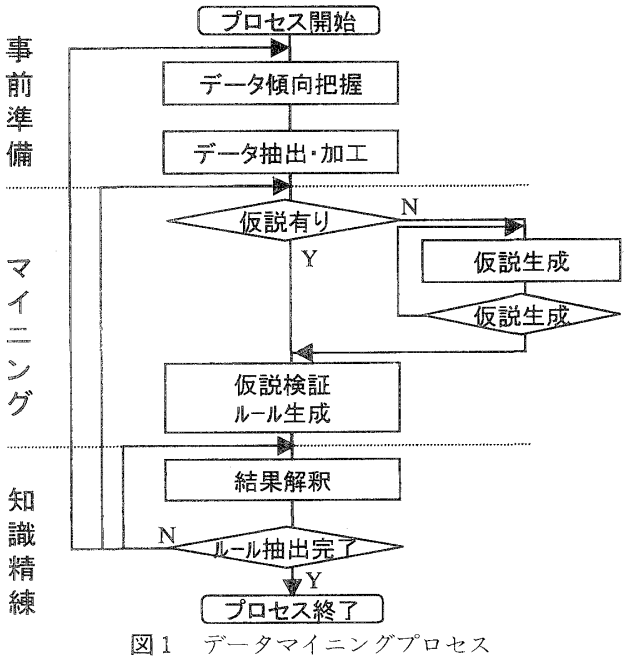


図１　データマイニングプロセス

下，各分類ステップ毎にその内容と特徴を述べ，要求される機能についてまとめる．

(1) 事前準備ステップ

本ステップは，前述の①アプリケーション領域の学習，②ターゲットデータ集合の生成，③データクリーニングと前処理，④データ削減と射影，⑤データマイニングの目標設定，⑥マイニングアルゴリズムが対応する．①のためには，業務の目的を明確にし対象データを特定する必要がある．②③④のためには，大量データのデータ型，値範囲，値分布等のデータ観察が必須である．データ型，欠損値，値範囲，データ件数等は，以降のアルゴリズム適用に影響する．⑤⑥は，分析経験に基づいてデータを観察し，人間が思考する作業であり，本ステップと次ステップの前半にわたる作業である．

本ステップには，マイニング戦略の立案という重要な役割があり，データ件数，データ型，値範囲，値分布，欠損値等のデータ観察やサンプリング，欠損値修復等のデータ加工支援機能が要求される．従来，データ観察には基礎統計量の数値表現が多用されていたが，数値データしか対応できない，平均的全体傾向しか分からない，統計解析の素人には直観的な理解が難しい等から，個々のデータを図形表現によって視覚化する機能と，簡易にデータ加工を支援できる機能の提供が有効と考えられる．

(2) マイニングステップ

本ステップは，前述の⑤データマイニングの目標設定，⑥マイニングアルゴリズムの選択，⑦マイニングアルゴリズムの実行が対応する．基本的には，大量データからのルール抽出と評価のための膨大な計算処理が対応し，現実的なマシン環境の中で適切な時間内に応答が得られる高速化の手法が求められる．

しかし，経験の少ない分野の分析や新しいルールの発見的分析では，事前準備ステップにおけるデータ観察だけではアルゴリズムの選択が難しい場合がある．このような場合，人間による試行錯誤の分析が必要であり，これを支援できる環境が求められる．１つの支援方法として，データ視覚化技術を応用して，データの特徴を効果的に自動視覚化し，人間の思考にヒントを与えることが考えられる．

(3) 知識精練ステップ

本ステップは，マイニング結果を解釈し，ビジネス上有用な知識を導出する過程であり，人間中心の作業である．意思決定の判断に使用するためには，外部環境や経営方針等の条件を考慮した高度で知的なフィルタリングを支援する環境が求められる．従来，知識精練には，フィルタリングを数式化しアルゴリズム処理させる方法と，人間がマイニング結果から有効なルールを選別する方法があった．

しかし，不要なルール除去等のフィルタリング処理を数式表現するためには，日々変更される背景知識や環境条件等の数値表現やデータ更新管理が求められ，多くの困難が想定される．そこで，数式化が難しい分野のフィルタリングを支援するためには，アルゴリズムが候補ルールを抽出し，そこから人間が知識精練を行うといった人間とマシンの協調作業が重要になる．１つの方法として，データ視覚化技術を応用して，ルールの構造や信頼度，関連情報との関係等を視覚化して，人間のフィルタリングを支援することが考えられる．

以上述べたように，データマイニングでは人間とマシンの協調関係が重要であり，この協調関係を支援するためには，人間とマシン間のインタフェースを司る視覚化技術の適用が鍵となる．以下，視覚化技術を適用した人間・マシン協調型データマイニングシステムの構成法について述べる．

3．人間・マシン協調型マイニングシステムの構成

3.1 視覚化技術の現状

本システム実現の鍵となる視覚化技術の現状について述べる．

(1) プロセス定義の視覚化：OLAP ツールに見られるスプレッドシートを利用した次元変更操作インタフェースが代表的である．最近は，プロセスの設計から定義・実行までの一連の機能と，自由な試行錯誤が簡易操作できるように，プロセスの単位機能と実行順序をグラフィカルに定義できる画面インタフェースが製品化されつつある．

(2) 源データの視覚化：図形の配置と色で多次元データの比較を容易化したピクセルマッピング[2]，源データの値範囲や値分布を可視化したパラレルコーディネート[3][4]，多次元データを図形の形状や配置に写像し２次元／３次元表現する多次元図形表現[5][6]等の多くの方法がある．これらは，事前準備ステップのデータ観察，OLAP 的な多次元データ分析，簡単なルールの提示等に使用されており，図形表現方法の変更，表示対象範囲の変更，座標変換等の視点変更イ

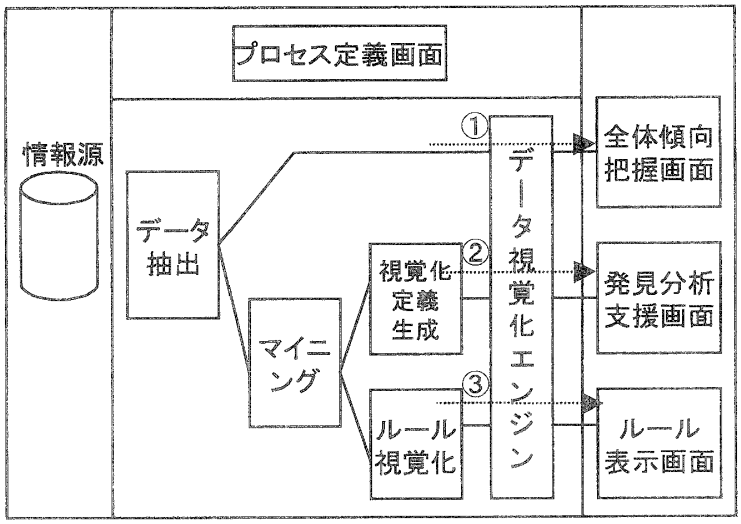


図2　データマイニングシステムの構成例

ンタフェースの高度化が検討されている。

(3) ルールの視覚化：決定木をツリー図表現した例や，相関ルールをネットワーク図等で表現した例[7][8]がある。しかし，ルール視覚化に関する研究は少ない現状にある。

3.2　システムの構成法

人間とマシンがお互いの長所を活かして，試行錯誤的なデータマイニングプロセスを総合的に支援する人間・マシン協調型データマイニングシステムの構成例を図2に示す。本システムは，分析対象データの抽出やデータクリーニングのためのデータ抽出，データの特徴やルールを導出するマイニング，発見的分析を支援する視覚化定義生成，知識精練を支援するルール視覚化，視覚化定義生成に基づきデータを視覚化するデータ視覚化エンジン，これら機能を統合的に操作可能なプロセス定義画面，および結果表示画面の要素機能から構成される。また，結果表示画面は，事前準備ステップを支援する全体傾向把握画面，マイニングステップの前段を支援する発見分析支援画面，知識精練ステップを支援するルール表示画面から構成される。ここで，データ視覚化エンジンには，データの意味・内容に依存しない多様な多次元表現が，外部制御によって簡易に得られる必要があり，この要求を満足するシステムとして視覚的多次元データ分析システム（INFOVISER）を採用した[9][10]。

全体傾向把握とルール視覚化は，表現方法の高度化がキーとなり，データの意味・関係を効果的に表す図形の形状表現や配置方法が求められる。一方，発見的分析支援のためには，マイニングアルゴリズムとデータ視覚化の連携がキーとなる。この具体例としては，決定木や相関係数のアルゴリズムから，目的属性の判別に因果関係の高い属性グループや相互に相関の高い属性グループを抽出し，各属性を視覚化の軸（配置や色，大きさ等）に反映させて自動的に視覚化し，パターン発見の候補となる画面を提示することにより，人間の思考にヒントを与えて発見的な分析を支援することができると考えられる[11][12]。

4．健康診断データのマイニング適用例

表1のような健康診断データを用いて，パターンを発見するプロセスの例を視覚化表現の有効性の観点から示す。

4.1　データの傾向把握

データの全体傾向把握（図2①）のためには，散布図マトリックス，ピクセルマッピング，パラレルコーディネート等の視覚化手法がある。散布図マトリックス（図3-1）は，図形の配置と形状で多次元の属性項目間の相関関係を把握しやすい。ピクセルマッピング（図3-2）は，指定された属性項目を基準にソートし，各属性項目の値を色のグラデーションで表現することにより，属性項目間で強い関係を持つ値の範囲が発見しやすい。パラレルコーディネート（図3-3）は，各属性項目ごとの値の範囲や分布と隣り合う属性項目間の値の共起関係が分かりやすく，欠損値の把握等にも有効である。

この例では，図3-1からγGTPは総コレステロールと尿酸の大小にかかわらず値が低いという傾向が，図3-2から尿酸が高いほど，γGTPが高い傾向が見られる。図3-3からも同様の傾向が見られる。このような全体傾向一覧表現は，マイニングを始める前に

表1　健康診断データの構成

氏名コード	性別	基準年齢	身長	体重	標準体重比	血圧MAX	血圧MIN	総コレステロール	中性脂肪	HDL	ブドウ糖	尿酸	GOT	GPT	γGTP
12340000	1	30	166	61	98	105	63	193	81	72	93	5.2	16	18	5
12340001	1	25	178	80	116	131	84	160	80	46	91	7.1	22	29	19
12340002	1	25	178	61	88	114	62	119	79	54	88	6	17	10	4

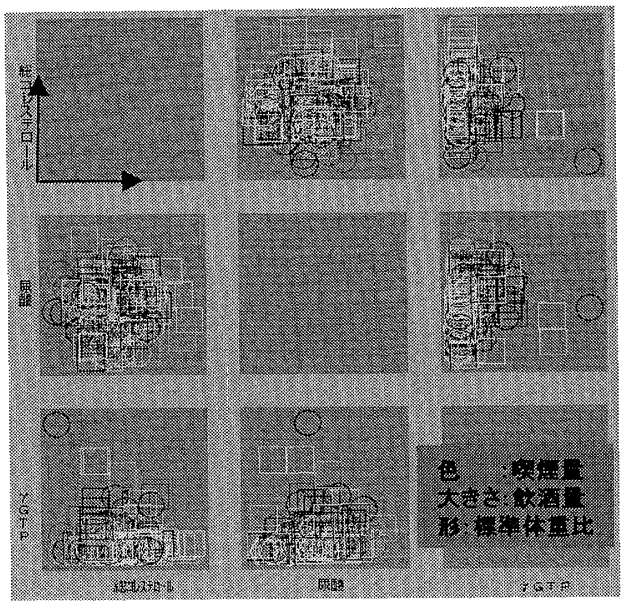


図3－1　散布図マトリックス

データの傾向を把握でき，プロセスの設計や結果解釈の支援に役立つと想定される．

4.2　発見的分析支援

発見的分析支援（図2②）のためには，決定木と相関係数を使用する2つの方法がある．目的属性がある場合は決定木から，ない場合は相関係数行列から関連性の高い属性グループを軸属性として抽出し，散布図を自動表示することができる．**図4**－(a)，(b)，(c)は，相関係数行列から，3つの手法（詳細は[11]参照）で視覚化属性を抽出して表示した結果を示す．この方法は，属性項目間の関連等のデータに対する知識が薄く，分析の進め方を決められない場合に有効である．

この例では，(a)，(b)は医学的知識の少ない人には直に抽出が難しい属性項目が軸として視覚化され，(c)は体重，身長，性別に強い相関が高いという比較的素人でも容易に想定できる当たり前のパターンが視覚化された．このように本支援機能は，あくまでルール発見の候補の提示であるが，分析者の思考にヒントを与えるという意味で有効と考えられる．

4.3　ルール視覚化

ルール視覚化（図2③）のためには，ルールごとにルールの持つ意味を効果的に表現できる方法がある．**図5－1**は決定木の視覚化例であり，エラー率や分岐数等を色や大きさ（幅）等で表現しているため，ツリー中のルートの重要度を直観的に把握できる．**図5－2**は相関ルールの視覚化例であり，共起関係を包含図

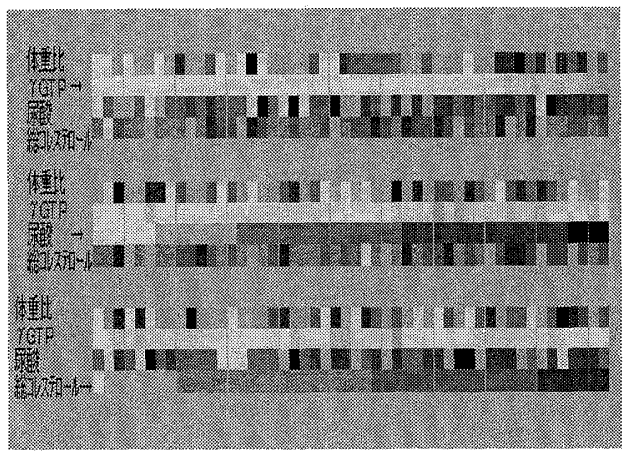


図3－2　ピクセルマッピング

で表すことにより，ルールが多い時やパスが2以上のルールがある時でも，全体傾向を把握しやすい．

図5－1は性別を目的属性とした決定木の視覚化結果で，エラー率でノードの色を，信頼度でラインの幅を表現しており，GPTとγGTPが一定値以上の場合に男性が多いことがわかる．図5－2は，各データ項目の値をカテゴライズして，分類間の同時出現頻度を相関ルールによって求めた結果を視覚化した例である（男性が出現するルールは除外した）．運動習慣が少ない場合に総コレステロールが高い傾向が見られる．このようなルール視覚化は，分析者の視点に基づくルールのフィルタリングに有効と思われる．

5．おわりに

本稿では，データマイニングにおける人間とマシン

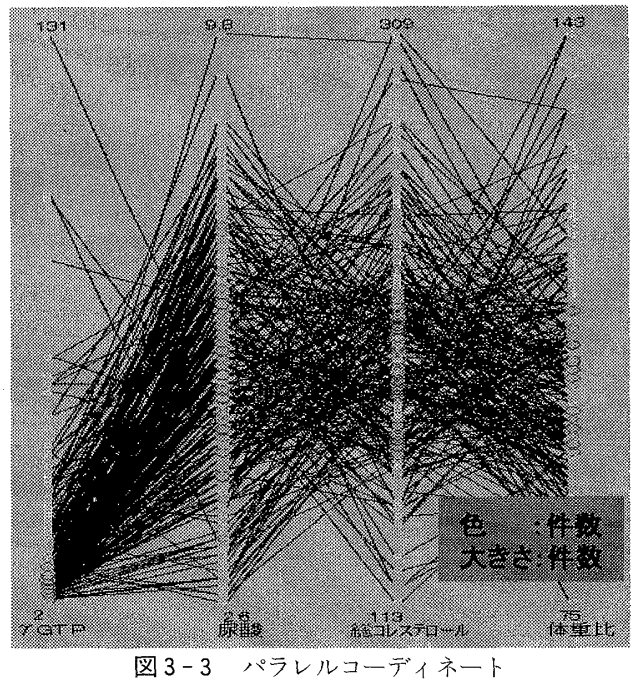
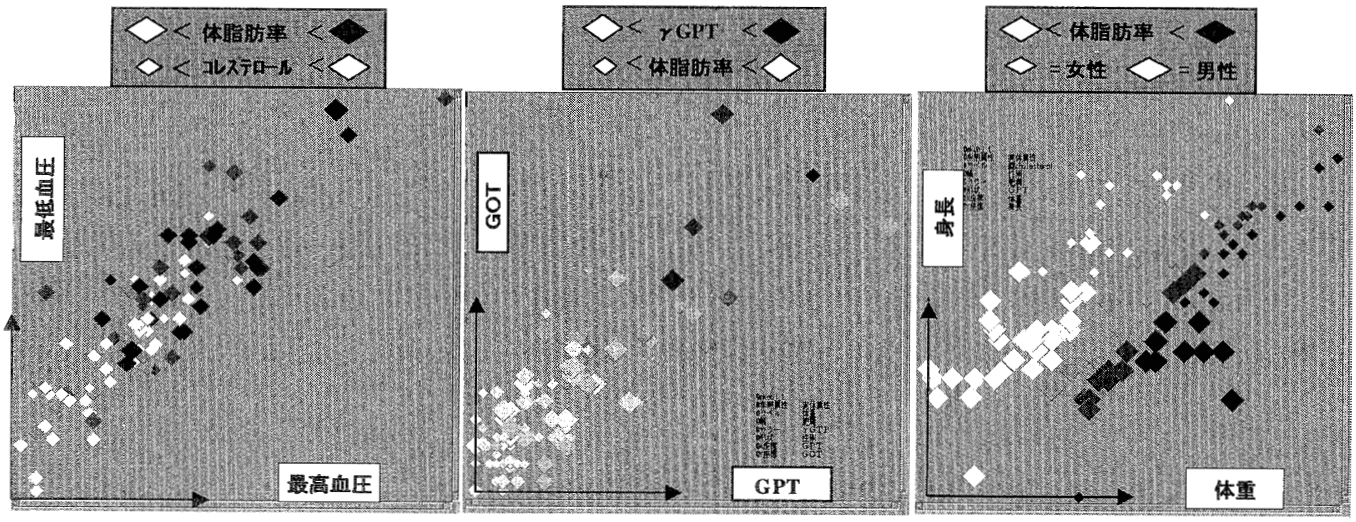


図3－3　パラレルコーディネート



(a) (b) (c)

図4　相関係数行列による自動視覚化例

の協調作業の必要性を述べ，実現のためにデータ視覚化がキー技術となることを示した．また，その要求に応える人間・マシン協調型データマイニングシステムの構成法を提案し，健康診断データを例題とした適用結果について示した．

最近は，受発注業務のような定型の基幹業務にもマイニングアルゴリズムを組み込んで，戦略的在庫管理を行うような新しいアプローチも現れている．一方では，これまで未経験の分野で新しいルールの発見を行おうとする挑戦的アプローチもある．本稿で提案したアプローチは，前者のようなアプローチには貢献できないが，後者のようなアプローチには役立つと考えている．

参考文献

[1] Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth「The KDD Process for Extracting Useful Knowledge from Volumes of Data」, Communications of the ACM, Vol.39, No.11, 1996.

[2] D.A. Keim, H.P. Kriegel「Issues in Visualizing Large Databases」, Visual Database Systems,

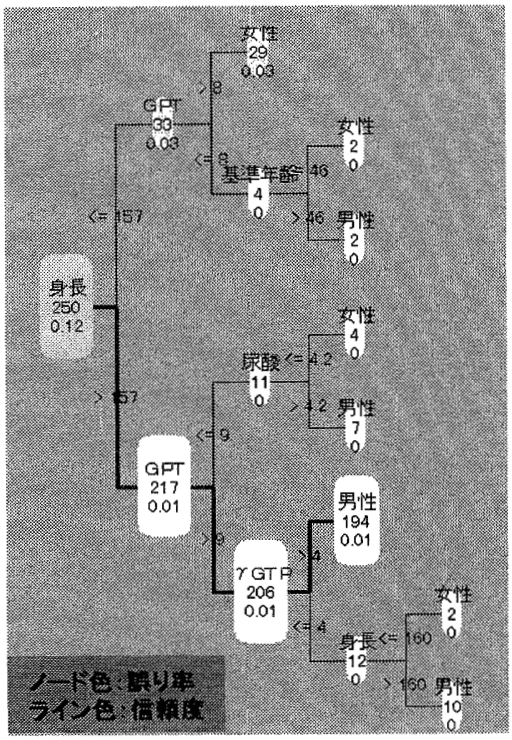


図5－1　決定木ルールの視覚化

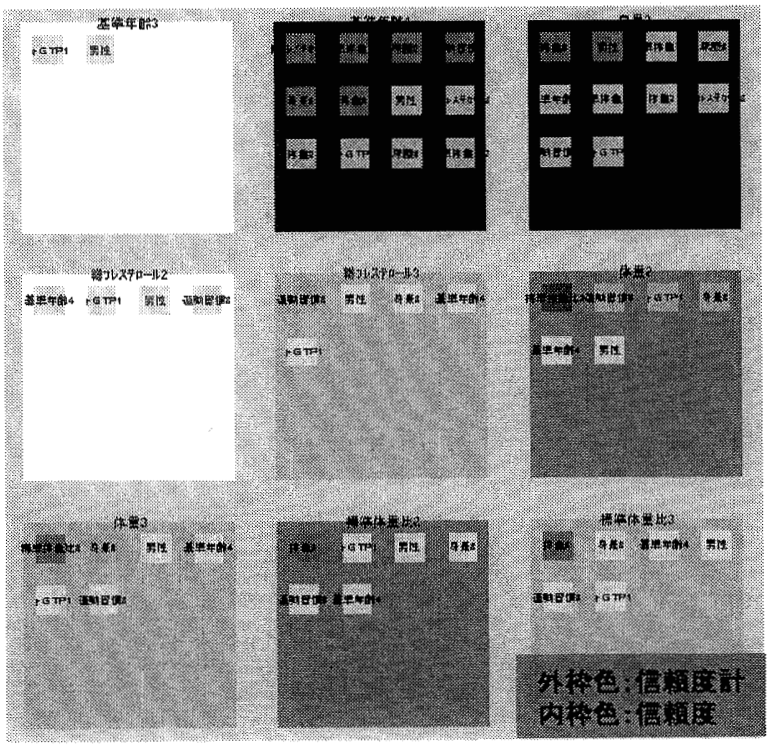


図5－2　相関ルールの視覚化

Chpman & Hall, 1995.
[3] hing-Yan Lee and Hwee-Leng Ong「Visualization Support for Data Mining」, IEEE EXPERT, 1996.
[4] Alfred Inselberg and Bernard Dimsdale「Parallel Coordinate：A Tool for Visualizing Multi-Dimensional Geometry」, IEEE Visualization, 1990.
[5] D.A. Keim,「Visual Data Mining」, The 23rd International Conference on Very Large Data Base (VLDB'97), Tutorial4, 1997.
[6] 増井俊之,「情報視覚化の最近の動向」, 電子情報通信学会, DEWS'98, チュートリアル, 1998.
[7] 福田剛志, 森下真一,「相関ルールの可視化について」, 電子情報通信学会, DE95-6, 1995.
[8] 飯塚哲也, 松尾比呂志, 磯部成二,「相関ルール視覚化によるデータマイニング支援方式」, 電子情報通信学会, DE97-68, 1997.
[9] 石垣昭一郎, 磯部成二,「データベースのビジュアル化ツール」, 経営システム, Vol.5, No.3・4, 1995.
[10] 磯部成二, 黒川　清, 塩原寿子,「DB情報ビジュアル化技術」, NTTR&D, Vol.45, No.1, 1996.
[11] 飯塚裕一, 飯塚哲也, 磯部成二, 梶原史雄,「相関係数行列の利用による視覚化対象属性選択方式」, 情報処理学会, DBS-113, 1997.
[12] 塩原寿子, 飯塚裕一, 丸山　猛, 磯部成二,「複数手法の組合せによる視覚化対象属性選択方式」, 電子情報通信学会, DEWS'98, No.48, 1998.