データベースシステム 117－1
（1999. 1. 22）

# 視覚的結果解釈支援方式の検討

丸山 猛 塩原寿子 飯塚哲也 磯部成二
{maruyama,shiohara,tetsuya,isobe}@dq.isl.ntt.co.jp

ＮＴＴ情報通信研究所
〒239-0847 神奈川県横須賀市光の丘 1-1

**あらまし** 近年、データベースに格納された顧客・販売情報データを利用した迅速で戦略的な意思決定を行うための技術として、分析者に専門的知識を要求しない視覚的なデータマイニングが注目されている。しかし、視覚的なデータマイニングでは、システムが自動的に結果を提示するため、分析者が結果を十分に理解して分析が行えないことがある。本稿では、この問題の原因を追求し、それを解決するために、視覚化結果の出力理由や分析を行うための視覚的結果のパターン提示による解釈支援方式の検討結果を示した。

**キーワード** 情報視覚化、データマイニング、結果解釈

# A Study on Interpretation Support System of Visualization Result

**Takeshi Maruyama,Hisako Shiohara,Tetsuya Iizuka,Seiji Isobe**

**NTT Information and Communication Systems Laboratories**
**1-1 Hikarino-oka Yokosuka-Shi,Kanagawa 239-0847,Japan**

**Abstract** Recently, visual data mining that doesn't demand enduser's knowledge of the analysis technique is paid to attention as a technology for strategic decision making. But enduser can't analyze data enough by using visual data mining system, because it is difficult to understand how visualization result is generated by system. This paper proposes the support method of interpretation that explains why the result is presented by system and where the characteristic exists in visualization result.

**Keyword** Visualization, DataMining, Interpretation

## 1．はじめに

コンピュータ技術が発達し、顧客・販売情報データをデータベースに格納することが可能になってきた。企業では、格納されたデータを分析して有用な情報を引き出し、迅速で戦略的な意思決定を行うことが要望されている。

このような要望に対し、データマイニングが注目されている[1]。データマイニングはデータのパターンを分析者に自動提示する技術である。その要素技術は、統計演算や機械学習、ルール導出等がある。分析者は各技術に特化した出力結果を認識し、データのパターンを把握する。

一方、分析者に技術的知識を要求しない分析手法としてデータの視覚化がある[2]。この手法では、人間のパターン認識能力や分析者の業務経験等の背景知識が活用でき、迅速かつ直観的な分析ができる。

我々は、このような背景のもと、データの視覚化によりデータの傾向の把握を支援するシステム INFOVISER-MINE の研究開発をしている[3]。INFOVISER-MINE では、システムが分析対象のデータ属性と図形の表示属性とのマッピングを行い、視覚化結果をユーザに提示する。分析者は、提示された結果から発見的な分析を行う。INFOVISER-MINE では、自動的なマッピングを実行するのに、データマイニングで利用される技術を利用している。つまり、INFOVISER-MINE は視覚的なデータマイニングを実現していると考えることができる。

しかし、上記の手法に則った視覚的なデータマイニングでは、システムが自動的に表示属性のマッピングを行うため、分析者が結果を十分に理解して分析が行えないケースがある。また、出力結果に対し、分析者の背景知識が先行して大きく影響することで、自明な結果のみが認識され、視覚的なデータマイニングを実現できないケースがある。これは、提示された視覚化結果に対し、この視覚化結果が出力された理由や分析の手掛かりとなるパターンの有無を十分に把握していないことが問題と考えられる。本稿では、この問題を解決するために、視覚化結果の出力理由や分析を行うための視覚化結果のパターン提示による解釈支援方式の検討を行った。

## 2．視覚的データマイニングの問題点

データ分析において、分析技術の専門的知識を必要としない方式として、データの視覚化がある。この方式は、文字数値データの属性と図形の色・大きさ・形・配置位置等の表示属性とをマッピングし、その結果を表示する。分析者は、経験に基づいた仮説のもと、属性マッピングを行う。その視覚化結果を観察し、興味がある属性を取捨選択し、仮説を立て直し、視覚化結果を再び作成する。この過程を繰り返し行うことで、様々な観点から視覚的なデータ分析を行い、パターン把握を実現する。しかし、この方法では、分析者が仮説を打ち立てる必要があるので、パターン把握できる有効な仮説を打ち立てるまで同じような分析を繰り返す必要があり、迅速性が失われる。

そこで、属性マッピングを自動的に行い、仮説の発見を促すという視覚的なデータマイニングの研究が行われている。我々は、属性マッピングの自動化に相関係数や決定木等のデータマイニング技術を利用する方式を提案し、これに基づいた視覚的データマイニング支援システム INFOVISER-MINE の研究開発を行っている[3]。これにより、データマイニングアルゴリズムによって出力された結果を利用した視覚化結果が提示され、その結果から分析・パターンの把握を行うことが実現できる。

しかし、分析者の観点からみると、視覚化結果の提示だけでは、システムが打ち立てた仮説がどのようなものなのか把握できないということや自動的に出力された視覚化結果の分析の切り口が見つけ出せないということにより、パターンの把握ができないことがある。また、分析者の背景知識が先行して大きく影響することで、自明な結果のみが認識され、視覚的データマイニングを実現できないことがある。

本検討では、このような問題に対し、分析者にパターン把握を促すための視覚化結果の解釈支援方式の確立を目的とした。

## 3．結果解釈支援方式の検討

前述した既存手法の問題点と背景から、仮説の発見を支援するための視覚化結果の解釈支援では、以下を実現する方式の検討が必要となる。

① 視覚化結果の根拠提示方式
② 視覚化結果の特徴提示方式

ここで、①で表現している根拠とは、その視覚化結果を提示する際に、どのようなデータマイニングアルゴリズムによってどのように表示属性を決定したのか、ということを表現している。これにより、前述したシステムが打ち立てた仮説を分析者に提示することができると考えられる。

また、②で表現している特徴とは表示された図形によって得られる視覚的な傾向、例えば図形の密集等、を表現している。これにより、前述した視覚化結果の分析の切り口を分析者に提示することができると考えられる。以下、各々の検討結果を示す。

## 4．各方式の検討結果

### 4．1．視覚化結果の根拠提示方式

この節では、第3章で述べた視覚化結果根拠提示方式の検討結果を示す。システムが立てた仮説を分析者が理解するには、その処理過程を提示することで実現できると考えた。具体的には、システムが与えられたデータから分析結果を出力するまでの過程、1．データ属性抽出、2．データ属性選択、3．データ属性と表示属性間のマッピング方式、に関しての情報を提示する。そこで、この3つの各々の処理過程で必要となるものを抽出した。以下に各過程で必要と思われるものを示す。

1．対象となるデータの情報（データ数、属性数等）
2．手段として利用したアルゴリズム(手法、パラメータ等）と選択された属性
3．マッピングの方式及びその処理結果

### 4．2．視覚化結果の特徴提示方式

この節では、第3章で述べた視覚化結果の特徴提示方式の検討結果を示す。まず、提示する特徴を分類する。分類した特徴の概要を以下の表 4.1 に示し、各々の特徴の説明及びその取得方法について示す。

表 4.1　検討した特徴とその概要

| 特徴 | 特徴の概要 |
|---|---|
| 傾向 | 2属性間の関数的な関係 |
| 密集 | 特定の1属性または多属性間の分布の疎密 |
| 段差 | |
| 突出 | |
| 周期 | 多属性間の周期関係 |

#### 4．2．1．傾向

2つの表示属性についての関数的関係を示す特徴である。例として、配置位置（ＸＹ軸）について述べる。この場合、単純な関数的関係として、直線的な配置（Y=AX+B）がある。図 4.1 にそのイメージを示す。その他、双曲線的な配置（Y=A/X+B）や指数的な配置（$Y=Ae^{bx}+C$）等がある。この特徴からは、属性 A が大きくなると属性 B が大きくなる等のパターンの発見支援ができる。

傾向を取得するには、定量的にその傾向の度合いを出力できる相関係数や回帰関数を利用する。そして、その特徴の精度には、Y’ =F（X）（X：選択した1属性）でモデル化した値 Y’ と実際の図形表示属性値 Y との誤差分散や相関係数値及びその有意性検定等が利用できる。

#### 4．2．2．密集

同様な表示属性値をもつ図形が集中して存在する特徴である。例として、配置位置と形状について述べる。左下（ＸＹ軸値がともに小さい）に円形が集中している状態である。図 4.2 にそのイメージを示す。この特徴からは、複数属性間での主要な値のパターンの発見支援ができる。

密集を取得するには、図形の表示属性値を与えることで、表示属性値分布の疎密を取得できる集約処理を利用する。その結果から、図形が多く集約された中心点を検出し、密集しているという特徴を取得する。取得の際の精度に関しては、集約結果の図形数や全体図形数との比や集約の中心点と集約された図形の表示属性値の距離分散等を利用する。また、取得した密集を表現するための表示属性値は、各集約結果の中心点の値を利用する。この結果と表示画面の分割により生成した領域とを利用することで、配置位置を主軸とした密集を検出することができる。

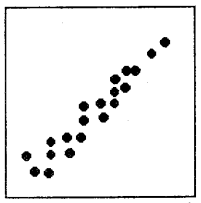
図4.1 傾向のイメージ

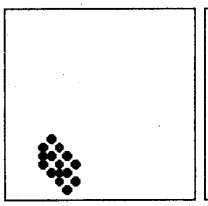
図4.2 密集のイメージ

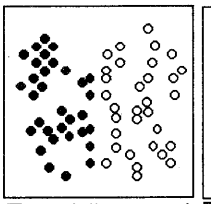
図4.3 段差のイメージ

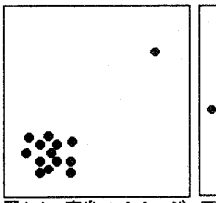
図4.4 突出のイメージ

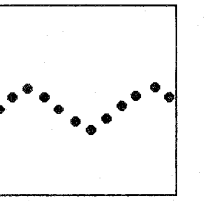
図4.5 周期のイメージ

### 4．2．3．段差

特定の表示属性をもつ図形に着目し、その図形がある表示属性値を境に他の表示属性値が異なっている特徴である。

例として、配置位置と形状について述べる。X座標の中点付近を境に左側は塗りつぶしの円が多く存在し、右側は中抜きの円が多く存在するという状態である。図 4.3 にそのイメージを示す。複数の属性による段差表現、例えば直線 Y=AX+B の上下での属性値の段差、も可能である。この特徴からは、特定属性のセグメンテーションパターンの発見支援ができる。

段差を取得するには、密集で利用した集約処理を利用する。これは、段差という特徴が、広義の密集という特徴としてとらえることができると考えられるからである。例えば、図 4.3 のイメージでは、左側には塗りつぶしの円の大きな密集が存在し、右側には中抜きの円の大きな密集が存在すると考えることができる。この考え方と密集の結果を段差取得に利用する。

まず、前述した密集と段差の関係から、密集で定義した分割数より少ない領域の作成を行う。次に、各領域に入る集約結果同士で、各々の表示属性値が共通している場合の数を調べる。各領域で上記計算を行い、その結果を比較する。比較の際の精度としては、各領域に入る該当表示属性値をとる図形数の比率や比較領域での図形数の比率等を利用する。

### 4．2．4．突出

特定の表示属性値を持つ図形が同様な表示属性値をもつ図形と比較し、特定の表示属性値だけが極端に違う特徴である。例として、配置位置と形状について述べる。左下、つまりＸＹ座標値がともに小さい、に円形が多く存在しているが、ただ１つの図形が右上、つまりＸＹ座標値がともに大きい、という状態である。

図 4.4 にイメージを示す。この特徴からは、異常値のパターンの発見支援ができる。

突出を取得するには、密集の取得方法を利用することができる。前述の定義から、突出は密集の補集合と考えることができる。よって、集約結果の中で、密集として取得できない集約結果から突出は取得できる。この取得する際の指標（突出の精度）としては、該当表示属性値をとる図形数や全図形数との比率等を利用する。

### 4．2．5．周期

ある特定の表示属性値が変化すると、それに対応して、他の表示属性値が周期的に変化する特徴である。例として、配置位置ついて述べる。X 座標の値を大きくすると、Y 座標の値が正弦値を繰り返す状態である。図 4.5 にそのイメージを示す。この特徴からは、周期パターンの発見支援ができる。

周期を取得するには、多変量解析や時系列解析で利用される正弦曲線近似法を利用する。近似法としては、フーリエ変換やスペクトル変換[4]等がある。これらの手法を適用することで、周期を表す周波数を取得する。但し、近似法は、通常１次元のデータに対し適用させる手法が多い。よって、本検討で利用するには、表示属性次元の削減を行う必要がある。配置された図形の傾向を保持しつつ次元削減を行うための KL 変換[5]等を利用することで上記は実現できる。また、取得の際の精度としては、取得した周波数の上限・下限等が利用できる。

## 5．適用性実験

### 5．1．実験概要

検討結果をもとに、適用性実験を行った。対象データは、健康診断データ（450 レコード,34 カラム）を利用した。

視覚化結果は、視覚的データマイニングを実現するツール INFOVISER-MINE による2次元の散布図結果とした。図 5.1 に INFOVISER-MINE が自動提示した視覚化結果を示す。自動提示したときに利用したデータマイニングアルゴリズムは相関係数とした。

検討した方式のうち、根拠提示方式及び特徴提示方式の傾向・段差・密集・突出の4特徴を本方式により計算した。傾向に関しては直線的な関係に限定した。この傾向の取得に Pearson の相関係数[6]を利用した。この手法は、相関係数の代表的手法であり、指標となる有意性が簡易に取得できる。他の特徴取得で利用する集約手法として、データの分布を損なわず高速に集約処理できる K-平均法[6]を採用した。

また、特徴表示の際に、取得結果の全てを出力するのではなく、特徴が顕著に現れている結果のみを選択した。そこで、出力する特徴の取得方法及び精度を表 5.1 のように設定した。

結果解釈の表示に関しては、HTML 言語及び WWW ブラウザを利用し、文字出力を行った。提示した各解釈結果を図 5.2,5.3 に示す。

表 5.1 各特徴取得方法

| 特徴 | 精度指標 | 値 |
|---|---|---|
| 傾向 | 相関係数値 | 0.75 以上 |
| | 優位性 | 99 %以上 |
| 段差 | 比較領域との図形数比 | 90 %以上 |
| 密集 | 全体図形数比 | 20% 以上 |
| | 表示属性値間の距離分散 | 平均距離以下 |
| 突出 | 全図形数との比率 | 1 %以下 |

## 5．2．考察

根拠提示方式に関しては、実験に利用した INFOVISER-MINE が出力する相関分析結果を検出して実現した。図 5.1 の結果の提示根拠を図 5.2 の結果で表示することで、システムが有効と判断した仮説の提示を実現している。これにより、分析者が仮説を把握し、背景知識を利用した属性修正等を行い、詳細な分析への移行ができる。

また、他の分析手法を適用する際には、以下の部分の文章を変更することで実現できる。

「表示属性を選択するのに必要なｱﾙｺﾞﾘｽﾞﾑ」

「表示属性に選択された属性」

その際に、利用したデータマイニングアルゴリズムの処理結果を取得し、結果を生成するインターフェイスが必要になる。また、根拠提示方式の表示に関して、文字を利用して出力を行った。データ属性数により、文字表示のみでは、煩雑になりやすい。根拠提示においては、視覚化結果の凡例とマッピングという点で密接しているので、今後は、根拠提示と凡例を融合した視覚的な根拠提示が有効になると考えられる。

特徴提示方式で、提示した特徴から把握できるパターンとしては、大きさ・色の縦横の段差から、「体重が軽い人は一般的に最高・最低血圧値は低く、軽くない人は一概に最高・最低血圧値が高いとはいえない」ということがいえる。血圧の値は体重と大きく関係することが良く知られている。しかし、対象とした実データに関してはそうではないことが、この解釈提示によって明らかになった。これは、分析者の背景知識の先行による視覚的データマイニングの問題を解消していると考えられる。さらに、視覚化結果を見ただけでは、円形の図形の直線的な配置から、「γ ＧＴＰが低い人は最高・最低が比例している」と把握する傾向がある。しかし、密集の特徴から、「実データでは、γ ＧＴＰと体重が低い人は最高・最低血圧も低い」ということが提示されている。これにより分析者は、直観的なパターンだけでは捕らえにくい新たな分析の切り口を把握し、視覚化結果を利用した詳細な分析への移行が可能になる。

また、検討した特徴を出力する以外に、基本的な統計量（平均・分散等）を特徴として提示する方法もある。しかし、視覚化結果を利用する際には、「このあたりに平均がある」等の表現ができる提示結果が理想的と考えられる。今回、検討した特徴では、基本統計量は見えないが、密集や段差を利用した大まかな平均や分散の把握は可能である。さらに、指標の精度を分析者が動的に変化させて、取得できる特徴やその変化を把握して、分析を行うことで、より有効に解釈結果を利用できると考えられる。

## 6．まとめ

結果解釈支援方式の検討にあたり、視覚化結果を利用した迅速かつ効果的なデータ分析を実現するために、①視覚化結果の根拠提示方式 ②視覚化結果の特徴提示方式が有効と考え検討・

試作を行った。

①に関しては、データ入力から視覚化結果出力までを大きく3段階に分け、各段階の情報を表示することを提案した。これにより、分析者がシステムから提示した結果に対し、その提示根拠を考えることなく、分析が行えることを実現した。②に関しては、特徴を大きく5つに分け、各々の詳細を示し、各特徴の取得方法について提案した。これにより、視覚的データマイニングにおいて、分析者が視覚的結果を利用して、詳細な分析を行うための切り口を提示することが可能になった。

また、WWW 上の記述言語 HTML 言語を利用し、適用性実験を行い、実現性・有効性を明らかにした。

今後の課題としては、以下が挙げられる。

・分析者による結果解釈の主観的評価
・散布図以外での結果解釈の適用
・分析者とのインタラクティブ性（精度等）
・結果解釈結果の提示の多様化（解釈結果の背景表示等）

**＜参考文献＞**


[1]U.M.Fayyad, G.Piatetsky_Shapiro, P.Smyth, R. Uthurusamy: "Advances in Knowledge Discovery and Data Mining", AAAI/MIT Press, 1995.

[2]S.G.Eick,D.E.Fyock: "Visualization Corporate Data", Proceedings of 20th VLDB Conference Santiago, p.487-499, 1994.

[3]飯塚，黒川，磯部:"視覚的データマイニング支援のための仮説生成方式", 第8回データ工学ワークショップ(DEWS'97)，pp. 37-42，1997.

[4]http://lips.is.kochi-u.ac.jp/image/fourie.html

[5]上坂,尾関: "パターン認識と学習のアルゴリズム", 文一総合出版, p. 109-120，1990.

[6]田中,脇本: "多変量統計法", 現代数学社, 1983.


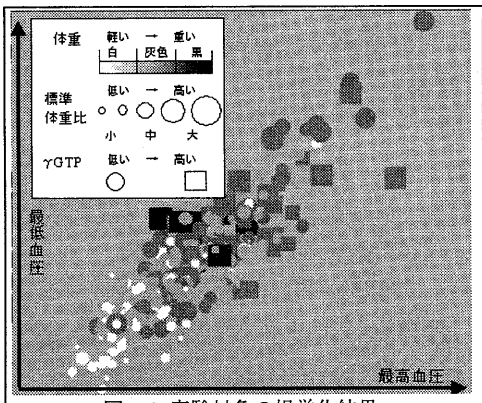


図 5.1 実験対象の視覚化結果

この視覚化結果を提示した理由

対象データ:医療データ(450 レコード,34 カラム)

視覚化対象データ属性候補

| | | | | |
|---|---|---|---|---|
| 性別 | 標準体重比 | 総コレステロール | 尿酸 | 飲酒回数 就寝前食事 |
| 基準年齢 | 最高血圧 | 中性脂肪 | GOT | 飲酒量 睡眠時間 |
| 身長 | 最低血圧 | HDL | GPT | 喫煙量 運動習慣 |
| 体重 | 降圧剤服用 | ぶどう糖 | γGTP | 食事回数 |

表示属性を選択するのに利用したアルゴリズム

全属性間の相関係数値を利用しました

表示属性に選択された属性

相関係数値が最も高い属性は**最高血圧**と**最低血圧**でした(相関係数値0.866)
この2属性を配置位置に対応させました
**血圧MAX**と**血圧MIN**の両方と**相関係数値が高い属性を上から順に表示属性**にしました

| | 対応表示属性 | 最高血圧との相関 | 最低血圧との相関 |
|---|---|---|---|
| 体重 | 色 | 0.284 | 0.28 |
| 標準体重比 | 大きさ | 0.264 | 0.254 |
| γGTP | 形状 | 0.232 | 0.263 |

図 5.2 視覚化結果の根拠提示

視覚化結果の特徴提示

**傾向**

配置位置Xと配置位置Yには正の相関関係があります(相関係数値0.875)
色と大きさには正の相関関係があります(相関係数値0.764)

**密集**

表示画面を縦横に4分割して図形の密集の有無を調べました

| | | |
|---|---|---|
| 領域2 | 領域1 | 領域3に灰色の中円が密集しています(148) 領域3に白色の小円が密集しています(202) |
| 領域3 | 領域4 | |

**段差**

表示画面を縦・横に中央で2分割して図形の段差の有無を調べました
・横方向の2分割
図形比率

| | | | |
|---|---|---|---|
| 93% | 7% | 色の段差があります | 大きさの段差があります |
| | | 白い図形は99%左側 | 小さい図形は99%左側 |
| | | 黒い図形は100%左側 | 大きい図形は97%左側 |

・縦方向の2分割
図形比率

| | | | |
|---|---|---|---|
| 5% | 色の段差があります | 形状の段差があります | 大きさの段差があります |
| 95% | 白い図形は99%下側 | 丸い図形は96%下側 | 大きい図形は97%下側 |
| | 黒い図形は100%左側 | | |

**突出**

表示画面を縦横に4分割して図形の突出の有無を調べました

| | | | | |
|---|---|---|---|---|
| 領域2 | 領域1 | 領域1に灰色の小矩形が突出しています(4) | 領域1に灰色の大円形が突出しています(1) | 領域1に灰色の大矩形が突出しています(4) 領域2に灰色の大円形が突出しています(3) |
| 領域3 | 領域4 | | | |

図 5.3 視覚化結果の特徴提示