データベースシステム 117－2
（1999. 1. 22）

データのカテゴライズと特徴に基づく視覚化形式の検討

飯塚裕一 塩原寿子 飯塚哲也 磯部成二
NTT 情報通信研究所
〒239-0847 神奈川県横須賀市光の丘 1-1
{iizuka, shiohara, tetsuya, isobe}@dq.isl.ntt.co.jp

大量データに内在する意思決定に有効なパターンやルールを発見し，提示する技術として，データマイニングが注目されている．しかし，複雑なアルゴリズムにより出力される結果は解釈が難しいという問題点があり，我々はユーザフレンドリな手法である視覚化を組み合わせた視覚的データマイニング支援の研究を進めてきた．これまで，マイニングアルゴリズムの適用により，データの特徴を効果的に表現すると思われる属性を自動的に選択し，その属性を図形の配置，形状に対応させることにより多次元散布図を提示する方式を検討してきた．本稿では，データの全体的／部分的な特徴を把握するための視覚化形式として散布図，包含図について検討した．複数の視覚化形式で結果を提示し，分析目的に応じて選択することが迅速な分析を進める上で重要であると考えられる．

キーワード：視覚化，データマイニング，データベース

A Study on Data Visualization Based on Category and Characteristics

Yuichi Iizuka Hisako Shiohara Tetsuya Iizuka Seiji Isobe
NTT Information and Communication Systems Laboratories
1-1 Hikarinooka Yokosuka-Shi, Kanagawa 239-0847 Japan
{iizuka, shiohara, tetsuya, isobe}@dq.isl.ntt.co.jp

Data mining has been paid attention as a technique for introducing business success. However the interpretation of mining results seems to be difficult. Then we have been studied a visual data mining support system applied visualization technique. Up to now, we have examined an automatic attributes selection mechanism utilizing some mining algorithms. By mapping selected attributes to plot profile such as X-axis, Y-axis, shape, color, size, and label, multi-dimensional representation of a data have been provided. In this paper, we examine data visualization patterns to grasp partially characteristics of data.

Keywords: Visualization, Data Mining, Database

1 はじめに

市場競争が激化する中で，データベース等に格納された大量データを有効に活用し，ビジネスに役立てることが，企業における重要課題となっている．このような状況の中で，大量データに内在するパターンやルールを抽出するデータマイニングが注目されている[1,2]．しかしながら，マイニングアルゴリズムは使用法や結果解釈が難しいため，専門的な知識なしには使いこなせないという問題がある．

我々は直観的にデータ傾向を把握するために有効な視覚化技術とデータマイニング技術を組み合わせた視覚的データマイニング支援方式の研究を進めてきた[3]．この方式は，特別な分析知識を持たないビジネス分野のエンドユーザが簡易に利用でき，業務知識や分析スキルを最大限に活用できると考えている．これまでに，データマイニング技術を分析対象データ全体または部分的データ中で有効と思われる属性の選択に利用し，選択された属性を，視覚化定義に利用する方式を検討してきた[4-7]．視覚化形式としては，選択された属性を図形の配置，形状(色，大きさ，形，ラベル)にマッピングして多次元的な関連性を散布図として提示している．

本稿では，データの全体的／部分的な特徴抽出に基づいて半自動的に視覚化提示する際に，部分的なデータ特徴の把握や部分データ間の特徴の比較を容易に行うための視覚化形式について検討した結果について述べる．

2 視覚的データマイニング支援方式

視覚的データマイニング支援方式は，システムからユーザへの情報提示方法として視覚化を適用したユーザ介在型のデータマイニング方式ととらえられる．この方式は，視覚化提示されたデータから人間が持つパターン認識能力や背景知識を利用して有効なパターンを発見するプロセスをシステムにより支援するものである．

我々は，研究開発を進めてきた"視覚的多次元データ分析ツール(INFOVISER)"を核として，視覚化技術により知識発見の各プロセス(データクリーニング，データマイニング，知識精練等)を支援する環境の構築を目指している．INFOVISER では，個々のデータと図形を１対１に対応させ，データ属性を図形の配置，形状(色，大きさ，形，ラベル)にマッピングしてデータの多次元性を表現することができる[8]．これにより，複数属性間の関連性の把握と個々のデータの確認を容易に行うことができる．視覚的データマイニング支援としては，決定木や統計演算，相関ルール導出アルゴリズムなどを利用して，分析対象のデータから特徴的な属性を抽出し，視覚化定義を自動的に生成する方式を検討してきた．

3 カテゴライズと属性選択および従来の視覚化

データ分析の際に，全体的な傾向のみでなく部分的なデータが持つ特徴を把握することが重要となる場合があることから，データの特徴に基づいた属性選択に加えて，データを部分集合に分割するカテゴライズについて検討してきた[6]．以下に，これまでに検討してきたカテゴライズ，属性選択方式，視覚化について簡単に述べる．

3.1 データのカテゴライズ

データのカテゴライズ手法として検討した方法を以下に示す．

- ディスクリミネーション
  指定される一属性の値による分割に有効
- クラスタリング
  指定される複数の属性の値による分割に有効
- 因子分析
  分割目的の属性が不定の際に有効

3.2 視覚化対象属性選択方式

分析対象のデータの特徴抽出手法としては以下を検討してきた．

- 決定木
- 相関係数行列
- 基礎統計量
- 相関ルール

これらの手法を用いて特徴として得られる出力に基づいて，視覚化対象とする属性を選択する．たとえば，決定木を利用する場合，目的属性の値の決定要因を特徴としてとらえ，上位ノードから順に優先順を付与して属性を選択する．また，カテゴライズ

と組み合わせた手法では，他カテゴリと比較して特徴的なカテゴリを検出し，そのカテゴリを特徴付ける属性を選択する方式を検討した．

3.3 従来の自動的な視覚化

上述のカテゴライズおよび属性選択方式により，優先順を付与して選択された属性を，配置(X 軸，Y 軸)および形状(色，大きさ，形，ラベル)にマッピングすることにより多次元的な散布図を自動的に提示する方式を検討してきた．データの属性と図形配置，形状の対応付けを決定するマッピング方法については，データ値の種別(連続／離散)や属性のもつ役割(目的要素／説明的要素)などにより決定している．

4 カテゴライズおよび特徴に基づく視覚化

4.1 散布図による視覚化形式の問題点

2 次元散布図による多次元表示は，複数の属性を図形の配置，形状に対応させることで属性間の関連性や個々のデータの分布状況を把握できるため，データ特徴の把握に有効である．また，特徴的なカテゴリの検出に基づいて属性選択し，カテゴリを色や形により表示することで，検出されたカテゴリの特徴把握や他カテゴリとの比較ができる．

しかし，カテゴリの散布図表現では，カテゴリを図形の色等により表現するため，表示情報の多次元性が減少する．また，特徴的カテゴリの検出やそのカテゴリの特徴表示ではなく，ユーザが指定したカテゴリなど特定カテゴリの特徴を他カテゴリと比較したい場合や，全カテゴリの特徴を把握したい場合には，特徴的なカテゴリの検出よりはカテゴリ毎の特徴抽出および表示が必要と考えられる．カテゴリ毎の特徴抽出は，全データに対する処理をカテゴリ内のデータ集合毎に行うことにより可能である．

上述のように，散布図によるカテゴリ表現では，

・カテゴリの表現により生じる多次元性の減少
・特定カテゴリの特徴把握／比較が不充分
・全てのカテゴリの特徴把握が困難

などの課題がある．

4.2 本稿で提案する視覚化形式

上記の課題を解決するためには，各カテゴリに対して独立に，特徴抽出を行い視覚化することが必要と考えられる．そこで，カテゴリと個々のデータの関係を包含関係として表す視覚化形式を検討した．

表 1 カテゴリを表す図形の表示手法

	視覚化対象属性	配置，形状への対応付け					備考
		X 軸	Y 軸	色	形	ﾗﾍﾞﾙ	
手法 C-①	全データの特徴抽出に基づいて選択された属性	第 1 属性	第 2 属性	第 3 属性	目的属性	カテゴリ	目的属性が離散値の場合
手法 C-②		目的属性	第 1 属性	第 2 属性	第 3 属性	カテゴリ	目的属性が連続値の場合
手法 C-③	カテゴリを目的属性とした特徴抽出に基づいて選択された属性	第 1 属性	第 2 属性	第 3 属性	第 4 属性	カテゴリ	—

カテゴリの持つ属性値は，カテゴリに含まれるデータの平均値とする

表 2 個々のデータを表す図形の表示手法

	視覚化対象属性	配置，形状への対応付け					備考
		X 軸	Y 軸	色	形	ﾗﾍﾞﾙ	
手法 D-①	1 つのカテゴリの特徴抽出に基づいて選択された属性 *全カテゴリで共通	第 1 属性	第 2 属性	第 3 属性	目的属性	第 4 属性	目的属性が離散値の場合
手法 D-②		目的属性	第 1 属性	第 2 属性	第 3 属性	第 4 属性	目的属性が連続値の場合
手法 D-③	カテゴリ毎の特徴抽出に基づいて選択された属性 *カテゴリ毎に異なる	第 1 属性	第 2 属性	第 3 属性	目的属性	第 4 属性	目的属性が離散値の場合
手法 D-④		目的属性	第 1 属性	第 2 属性	第 3 属性	第 4 属性	目的属性が連続値の場合

包含図を生成するために以下の処理を行う．

1)カテゴライズ

2)全データ，カテゴリ毎のデータから特徴抽出

3)カテゴリを表す図形の表示手法の決定

4)個々のデータを表す図形の表示手法の決定

5)データから図形への情報変換と視覚化提示

上記1)，2)は，3.2および3.3に示した手法を用いて行う．上記3)，4)の検討結果を表1，表2に示す．これらの表示手法を組に合せて処理することで，複数の視覚化形式で結果提示することができる．

5 適用例

データの全体的／部分的な特徴に基づいて視覚化提示した例について以下で考察する．

サンプルデータは，レコード数205，属性数26の自動車データであり，車長，車幅，車重，燃料タイプ，車種，ドア数，馬力，エンジンサイズ，ホイルベース，価格などの属性を持つ．カテゴライズは車種の値による分類で行い，"セダン"，"ハッチバック"，"その他"の3つのカテゴリに分けた．特徴抽出には決定木を利用し，目的属性を高価／廉価の2値を持つ価格として生成された決定木の上位ノードから順に優先付けて属性選択した．視覚化の際には，価格をデータに対応する図形の形に対応させ，高価な場合は矩形，廉価な場合は円形で表現した．全データおよび各カテゴリに含まれるデータから，選択された属性を表3に示す．

5.1 全データの特徴に基づく視覚化(散布図)

表3の全データから選択された属性に基づいて視覚化した例を図1に示す．属性をX軸，Y軸，色，大きさ，形，ラベルの順に対応させている．

全データ(全車種)を分析対象とした場合には，円形の図形が左上方に存在し，大きさが小さいことから，価格が廉価な自動車は，車重やホイルベースが小さく市街地燃費が良いことが分かる．また，図形の配置状況から，車重(X軸)と市街地燃費(Y軸)には反比例の関係があること，図形色の傾向から，車重(X軸)が同程度であれば圧縮比(色)が大きいほど市街地燃費(Y軸)が良いことなどが分かる．

このように，視覚化することで，自動車の価格に影響する要因(属性名)のみでなく，価格に影響する複数の属性間の関係や，個々の自動車が持つ属性値の分布状態を把握することができる．

図2に散布図によりカテゴリを表現した結果を示す．図1において圧縮比に対応させていた色に，カテゴリ(車種)を対応させて表現している．色が濃い図形は他と比較して小さく，左上方に配置され，色が薄い図形は右下に配置される傾向にある．また，配置，大きさに対応する各属性値の取っている範囲は異なるものの，いずれの色の図形も同様な傾向を

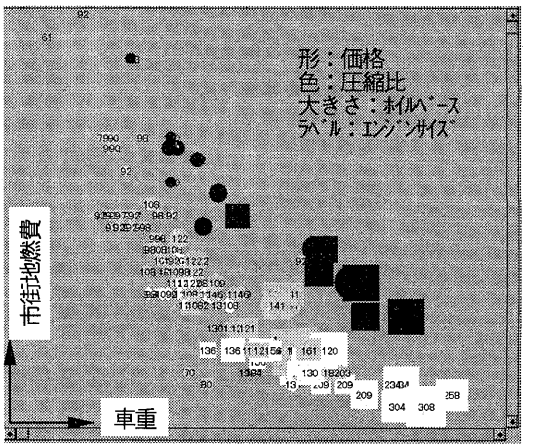

図1 全データの特徴に基づく視覚化結果

表3 全データおよび3つのカテゴリから決定木を利用して選択された属性

	全ﾃﾞｰﾀ(全車種)	カテゴリ		
		ｾﾀﾞﾝ	ﾊｯﾁﾊﾞｯｸ	その他
第1属性	車重	ﾎｲﾙﾍﾞｰｽ	車長	ｴﾝｼﾞﾝｻｲｽﾞ
第2属性	市街地燃費	馬力	馬力	車長
第3属性	圧縮比	ﾘｽｸ等級	ﾎｲﾙﾍﾞｰｽ	車高
第4属性	ﾎｲﾙﾍﾞｰｽ	車重	—	ﾘｽｸ等級
第5属性	ｴﾝｼﾞﾝｻｲｽﾞ	—	—	—

ﾘｽｸ等級：価格に対する保険ﾘｽｸを6段階評価したもの

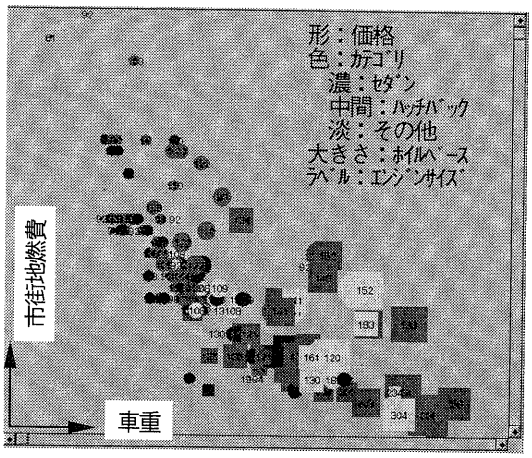

図2 全データの視覚化結果(色：カテゴリ)

示している．散布図による表現では，カテゴリ間の比較を容易に行うことができると思われる．

5.2 一つのカテゴリの特徴に基づく視覚化

カテゴリ間の特徴の比較は，図2に示した散布図で可能であるが，特定カテゴリのみのデータを見たい場合や，データ数が多く図形が重畳表示される場合には，カテゴリ毎に視覚化する包含図が有効であると考えられる．また，カテゴリを図形の色や形により表現する必要がないため，多次元性を保持することができる．

カテゴリと内包されるデータの関係および特定カテゴリの特徴を同時に提示するため，表1のC-①，表2のD-①の手法により包含図で提示する視覚化形式を検討した．カテゴリ数が3であるため，各カテゴリから選択された属性を個々のデータの配置，形状に対応させることにより，3つの包含図が提示される．図3は，"ハッチバック"のデータから選択された属性で全カテゴリに内包される各データの配置，色を決定した例である．カテゴリを表す図形は，全データから選択された車重，市街地燃費，圧縮比，価格で配置，色，形をそれぞれ決定している．

図3から，他カテゴリと比較した"ハッチバック"の特徴把握ができる．たとえば，各カテゴリを表す図形の配置，色から，"ハッチバック"は，他カテゴリと比較して平均的に，車重が軽く市街地燃費が良好であり，圧縮比が低い傾向にあることが分かる．また，"ハッチバック"の特徴に基づいて各カテゴリ内の図形の配置，色を決定しているため，"ハッチバック"で価格に影響する属性(車長，馬力，ホイルベース)が他カテゴリ内でどのような傾向を示すのかを見ることができる．"セダン"の場合には，"ハッチバック"と類似した傾向を示すのに対して，"その他"のカテゴリでは，車長や馬力の価格に対する影響は少なく，車長と馬力の比例的な相関が低いことが分かる．

5.2 カテゴリ毎の特徴に基づく視覚化

全てのカテゴリについて，カテゴリに内包されるデータの詳細な特徴を把握するためには，カテゴリ毎に特徴抽出を行い，選択された属性を視覚化する必要がある．そこで，表1のC-①，表2のD-③の

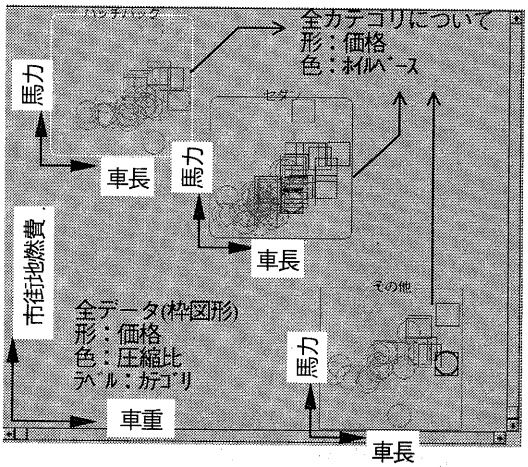

図3 ハッチバックの特徴に基づく視覚化結果

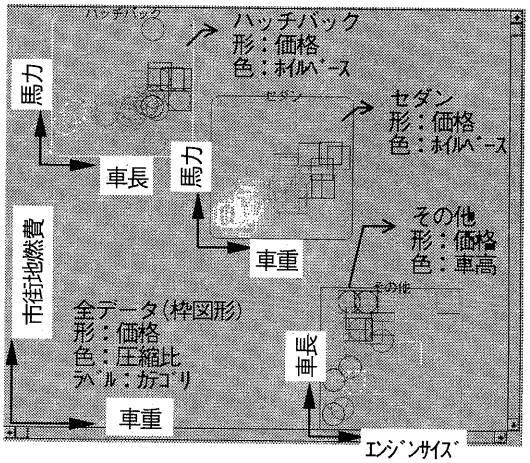

図4 各カテゴリの特徴に基づく視覚化結果

手法により包含図を提示する視覚化形式を検討した．図4は，図3と同様にカテゴリを表す図形の配置，色，形を決定し，各カテゴリから選択された属性で該当するカテゴリに内包される各データの配置，色を決定することにより視覚化した例である．

図3と同様に，他カテゴリと比較した"ハッチバック"の特徴(車重，市街地燃費，圧縮比)の把握ができる．また，カテゴリ毎に内包される個々のデータの配置，形状を見ることで，各カテゴリにおいて価格に影響する要因が分かる．たとえば，"ハッチバック"のみのデータ傾向は全データとは異なり，価格には車長，馬力，ホイルベースが影響し，特に馬力(Y 軸)が価格決定要因として大きいことが分かる．

6 考察

全体的な傾向や，特定カテゴリに特有な特徴の比較は，散布図による視覚化形式が有効である．

図3，図4で示した包含図による視覚化形式は，目的属性や全データを特徴付ける属性に対して影響が少ない属性でカテゴライズした場合に，カテゴリ毎の特徴を把握するのに有効であると考えられる．特に値が名義尺度の離散値などであり値間の関連がない属性で生成したカテゴリに注目した分析に役立つと考えられる．

図3の視覚化形式は，特定カテゴリに注目し，全データおよび特定カテゴリの特徴を合わせて表示できるため，"ハッチバック車で廉価な自動車の特徴が知りたい，また他車種と比較したい"のような要求に応えることができると考えられる．

図4の視覚化形式では，各カテゴリにおける特徴を提示することができるため，"全車の傾向および車種別の特徴を知りたい"のような要求に応えることができると思われる．

このように，全データに対するカテゴリの特徴把握および比較には散布図が有効であり，カテゴリ毎に異なる特徴把握や特定カテゴリに注目した特徴把握には包含図が有効であると考えられる．データ件数や視覚化に利用した属性の値分布などに依存する認識容易性を考慮すると，複数の視覚化形式をユーザに提示し，選択的かつ動的に目的とする結果を得られるようにすることが迅速な分析を実現するために重要であると思われる．

今後は，分析目的とする属性や全データを特徴付ける属性とカテゴライズする属性の関連性や，分析対象データの値の分布と視覚化結果の認識容易性との関係などを，種々のデータを利用して検証することにより，有効な分析を行うための視覚化方式を検討する．また，データ特徴やデータ間の関連性に応じて，樹系図や網状図などの視覚化形式も検討することが必要と考えられる．

7 おわりに

データのカテゴライズとおよび全データ／カテゴリ毎のデータの特徴に基づいて散布図，包含図により表現する視覚化形式について検討した．複数の視覚化形式による結果を提示し，カテゴリ間の比較や特定カテゴリ内のデータ特徴の把握などの分析目的に応じて選択することが，有効かつ迅速な分析を行う上で重要である．

今後は，考察に述べた課題について検討することで有効な視覚化形式の提示を目指す．

<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]河野，西尾，J.Han: "データベースからの知識獲得技術", 人工知能学会誌, Vol.10, No.1, pp.38-44, 1995.

[3]黒川，飯塚，磯部: "視覚的データマイニング支援方式の検討", 情報処理学会研究報告 96-DBS-110, Vol.96, No.103, pp.15-22, 1996.

[4]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.

[5]飯塚，飯塚，磯部，梶原: "相関係数行列の利用による視覚化対象属性選択方式", 情報処理学会研究報告, 97-DBS-113, Vol.97, No.64, pp.27-32, 1997.

[6]塩原，飯塚，丸山，磯部: "複数手法組合せによる視覚化対象属性選択方式", DEWS' 98, No.48, 1998.

[7]飯塚，磯部: "相関ルールを用いた視覚化属性選択方式", 信学技報, DE98-14, pp.9-17, 1998.

[8]黒川，磯部，塩原，鬼塚: "情報可視化のためのデータビジュアル化モデル", 情報処理学会研究報告 96-HI-65, Vol.96, No.21, pp.51-56, 1996.