データベースシステム 113－5
（1997. 7. 14）

相関係数行列の利用による視覚化対象属性選択方式

飯塚裕一† 飯塚哲也† 磯部成二† 梶原史雄‡

†NTT情報通信研究所
{iizuka, tetsuya, isobe}@dq.isl.ntt.co.jp

‡奈良先端科学技術大学院大学 情報工学研究科
masao-k@is.aist-nara.ac.jp

データベースに格納された大量のデータを有効活用し，ビジネスに役立てる技術としてデータマイニングが注目されている．データマイニングは，大量データに隠されたパターンやルールを発見する技術であり，関連分野から研究が進められている．我々は，人間の視覚による優れた認識能力に着目し，視覚的データマイニング支援方式について検討している．この方式では，データ特徴に基づいて視覚化対象属性を選択して結果提示する．選択された属性により得られるパターンが決定されるため，何をデータ特徴とし，特徴を属性選択でどのように利用するかが課題となる．本稿では，属性間の関連性を表す相関係数行列をデータ特徴とした場合の属性選択方式について述べる．

キーワード：データマイニング，視覚化，データベース，相関係数行列

Attribute Selection Method for Visualization with Correlation Coefficient Matrix

Yuichi Iizuka† Tetsuya Iizuka† Seiji Isobe† Masao Kajihara‡

†NTT Information and Communication Systems Laboratories
{iizuka, tetsuya, isobe}@dq.isl.ntt.co.jp

‡Graduate School of Information Science Nara Institute of Science and Technology
masao-k@is.aist-nara.ac.jp

Data mining has been paid attention as a technique for introducing business success. Data mining is to find effective patterns and rules from a large amount of data. Various approaches are attempted in some research fields. At the view point of harnessing the perceptual and cognitive capabilities of the human user, we focused on the visual data mining support system. On the visual data mining support, definition for visualization is important for obtaining the effective patterns. For the definition, how to find useful data attributes as a visualization target become key point. In this paper, we propose several methods of automatic selecting target attributes by using correlation coefficient matrix.

Keywords: Data Mining, Visualization, Database, Correlation Coefficient Matrix

1　はじめに

市場競争が激化する中で，データベースに格納された大量のデータを有効活用し，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．このような中で，大量データに内在するルールやパターンを抽出するデータマイニングが注目されている．データマイニングは，機械学習に関する成果をもとにデータベースからの知識発見（KDD: Knowledge Discovery in Databases）[1,2]として研究が進められている．KDDの主なプロセスは，ターゲットデータベース作成，データクリーニング，パターン抽出／検証，頑健化，知識化である．データマイニングの視点からKDDプロセスを見ると，ターゲットデータベース作成とデータクリーニングは前処理，パターン抽出／検証は本処理，頑健化と知識化は後処理と考えられる．

データマイニング手法として，判別やクラスタリング[3]，連想ルールの抽出[4]などがあげられが，これらの手法では結果が論理式や決定木で提示される．意思決定を行うためにデータマイニングを利用したいと思っているエンドユーザのためには，よりユーザフレンドリで柔軟なデータマイニング技術の確立が必要である．

我々は，データを視覚化表現し，直観的にパターンを発見する「視覚的データマイニング支援方式」について検討している[5,6]．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．

本稿では，「視覚的データマイニング支援」の重要な処理の一つである視覚化対象属性の選択を相関係数行列を用いて行う方式について述べる．

2　視覚的データマイニング支援

「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．その実現に際して，研究開発中の視覚的多次元データ分析ツール（INFOVISER）[7,8]の適用を検討している．INFOVISERでは，文字・数値データを図形情報に変換する定義（情報変換定義）をユーザが行うことで視覚化表現する．図形の配置や形状などにデータ属性を対応させることにより，多次元属性間の関連性を表現することができる．

一般に視覚化ツールを用いた分析では，ユーザが仮説をたてて視覚化のための定義を行い，結果の評価・検証と仮説修正を繰り返す．例えば相撲力士の特徴分析において，"力士の体格が強さに影響するであろう"という仮説をたてて，力士の身長をX軸，体重をY軸，地位を図形の大きさで表現する定義をユーザが行う．視覚化結果における，大きさが大きい図形（地位の高い力士を表す）の分布から地位の高い力士の体格の傾向が把握できる．

「視覚的データマイニング支援」では，システムがデータ特徴の抽出をもとに上述の身長，体重，地位のような属性を選択し，情報変換定義（図形表現への対応付け）を生成してユーザに視覚化結果を提示する．ユーザは，結果解釈を通じて情報変換定義を修正することで分析を進める．図2.1に「視覚的データマイニング支援」の処理

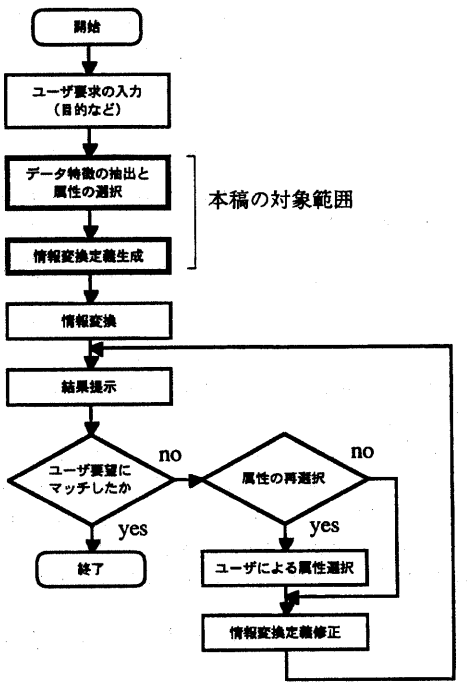

図2.1　視覚的データマイニング支援の処理フロー

フローを示す．ユーザの要求に応じてデータの特徴を抽出し，その特徴に基づいて情報変換定義を生成して視覚化結果をユーザに提示する．

3　属性選択方式の検討

データ特徴から属性を選択し，有用な視覚化表現を推定して提示する「視覚的データマイニング支援方式」では，何をデータ特徴として，どのように視覚化対象属性を選択するかが課題である．特徴抽出から定義生成までの検討方針および決定木を用いた場合の結果は文献[6]に述べた．

本稿では属性間の関連性をデータ特徴として着目し，相関係数行列の利用を検討した．相関係数は，２属性間の相関性を示す統計量である[9]．データがもつ多次元的な関連を有効表現する複数の視覚化対象属性を選択するために，２属性間の関連を如何に利用するかが課題となる．以下に相関の高さの捉え方として，

・２属性間での相関の高さ(最大相関係数)
・部分属性(≒視覚化対象)に関する相関の高さ
・全属性に関する相関の高さ

に着目した属性選択方式について述べる．属性A_iと属性A_jとの相関係数をC_{ij}で表す．

3．1　属性選択方式１

データに内在するパターンを見出すための前段階として，最大相関係数を有する属性対とその属性と相関が高い属性を選択し，データ特徴の一例として示すことを目的とする．最大相関係数をもとにした分析対象属性の絞込が可能になる．相関係数行列を導出した後の処理を以下に示す．

1) 最大相関係数C_{kl}を有する属性A_k,A_lを選択する．
2) 属性A_kと他属性との相関係数$C_{k1},C_{k2},\cdots,C_{kl},\cdots,C_{kn}$と，属性$A_l$と他属性との相関係数$C_{l1},C_{l2},\cdots,C_{lk},\cdots,C_{ln}$とを各々乗じ，$M_1,M_2,\cdots,M_n$（$M_i=C_{ki}\times C_{li}$）とする．
3) M_k,M_lを除いて，値の大きい方から順に抽出し，Mの算出に用いた属性を順次選択（M_iならばA_iを選択）する．

3．2　属性選択方式２

視覚化の最大表現次元数を考慮して，ユーザが指定した属性数の範囲で，相互に相関が高い属性群を選択し，データ特徴として示すことを目的とする．相関係数行列を導出した後の処理を以下に示す．

1) 選択属性数をmとし，各属性と他の属性との相関係数の中で大きい方から$m-1$個の相関係数を合計した値$S_1,S_2,\cdots,S_n$を求める．
2) 最大のS_kを有する属性A_kを選択する．
3) 属性A_kと相関が高い属性を順次選択する．

3．3　属性選択方式３

選択する属性数に制限を加えず，全属性に対して相関が高い属性を選択し，相関係数行列全体を代表する属性をデータ特徴として示すことを目的とする．相関係数行列を導出した後の処理を以下に示す．

1) 各属性について他属性との相関係数を合計した値$S_1,S_2,\cdots,S_n$（$S_i=C_{i1}+C_{i2}+\cdots+C_{in}$）を算出する．
2) 最大のS_kを有する属性A_kを選択する
3) 属性A_kと相関が高い属性を順次選択する

4　実験システム

相関係数行列を利用して視覚化対象属性を選択し，情報変換定義を生成，視覚化表示するシステムのプロトタイプを作成した．相関係数行列はSAS[10,11]を利用して導出した．

5　実施例

医療データを用いて，相関係数行列の利用による属性選択方式および得られた視覚化結果について考察した．

試験に用いたデータは，レコード数が103件であり，属性は"年齢"，"性別"，"身長"，"体重"，"血圧MAX"，"血圧MIN"，"総cholesterol"，"ブドウ糖"，"GOT"，"GPT"，"γGTP"，"肥満度"，"飲酒"，"喫煙"である．このデータから導出した，相関係数行列を表5.1に示す．また，各属性選択方

表5.1　医療データの相関係数行列

	年齢	性別	身長	体重	血圧MAX	血圧MIN	総cholesterol	ブドウ糖	GOT	GPT	γGTP	肥満度	飲酒	喫煙
年齢	1.000													
性別	0.094	1.000												
身長	-0.105	-0.742	1.000											
体重	0.022	-0.615	0.648	1.000										
血圧MAX	0.055	-0.043	-0.067	0.296	1.000									
血圧MIN	0.067	-0.224	0.041	0.358	0.879	1.000								
総cholesterol	-0.081	-0.180	0.096	0.446	0.443	0.440	1.000							
ブドウ糖	-0.057	-0.243	0.130	0.272	0.202	0.261	0.396	1.000						
GOT	-0.018	-0.308	0.128	0.368	0.283	0.285	0.204	0.083	1.000					
GPT	-0.019	-0.326	0.147	0.470	0.258	0.308	0.185	0.218	0.842	1.000				
γGTP	-0.078	-0.307	0.161	0.333	0.196	0.238	0.308	0.271	0.709	0.657	1.000			
肥満度	0.086	-0.230	0.063	0.743	0.504	0.512	0.600	0.268	0.397	0.485	0.334	1.000		
飲酒	-0.033	-0.589	0.446	0.469	0.042	0.248	0.214	0.201	0.169	0.214	0.242	0.265	1.000	
喫煙	-0.137	-0.557	0.469	0.372	0.073	0.177	0.165	0.132	0.212	0.203	0.238	0.197	0.465	1.000

表5.2　各方式で選択された属性

	属性選択方式1	属性選択方式2	属性選択方式3
第1属性	血圧MAX	GPT	体重
第2属性	血圧MIN	GOT	肥満度
第3属性	肥満度	γGTP	身長
第4属性	総cholesterol	肥満度	性別
第5属性	体重	体重	GPT

式により得られた属性群を表5.2に示す．

各属性選択方式の適用により得られた属性を用いて視覚化した結果を以下に述べる．基本的に，第1属性から順にX軸，Y軸，色，大きさに対応させている．

5．1　属性選択方式1の適用結果

表5.2の属性選択方式1の欄に示した属性を用いた視覚化結果を図5.1に示す．配置，色，大きさに対応付けられた複数属性について，値の分布状態（相関の高さ）の把握が可能である．全体傾向としては，"最高血圧と最低血圧は相関が強く（X軸とY軸の値が直線関係），血圧が高い人は肥満度が高い（右上に配置された図形は色が濃い）が，総cholesterolの相関は高くない（同じ大きさの図形が均一に分布）"ことが一瞥して分かる．また，図形毎に複数の属性を追って見ることができる．例えば，"血圧MAX，血圧MINがとも

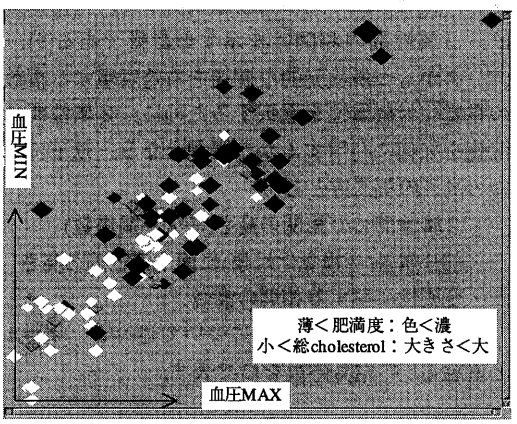

図5.1　属性選択方式1を適用した視覚化結果

に高い（X軸，Y軸ともに値が大きい）にもかかわらず，肥満度と総cholesterolが低い人（図5.1の中央の色が薄く，大きさが小さい図形）が存在する"ことが分かる．このように複数属性に関する全体傾向に加えて特異点などの把握ができ，これをもとに次段階の分析を進めることができる．

5．2　属性選択方式2の適用結果

表5.2の属性選択方式2の欄に示した属性を用いた視覚化結果を図5.2に示す．配置，色，大きさの分布状態から，"GTP，GOT，γGTPは相互に相関が高く，肥満度が低い人ではGPT，GOT，γGTPの各値が小さい"ことが分かる．特に，配置や色が連続的に変化する全体傾向に対して，大

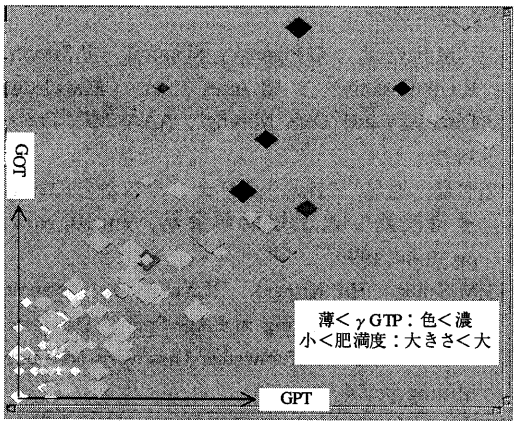

図5.2　属性選択方式２を適用した視覚化結果

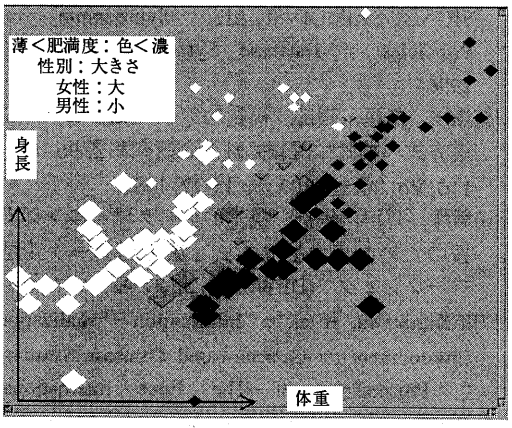

図5.3　属性選択手法３を適用した視覚化結果

きさに注目すると図5.2の左下に小さい図形（肥満度が低い人）が密集して存在する様子が分かる．また"γGTPの値が大きい（色が濃い）人は，GPT，GOT値がともに大きいだけでなく，特にGOT値が大きい（図中の右上に配置）"ことがわかる．以上のような傾向把握は，さらに詳細な分析を進める上で，"γGTPの大きい人（色が濃い図形）に着目して分析したい"というようなデータの絞込や分析方針の立案に利用できると考えられる．

５．３　属性選択方式３の適用結果

表5.2の属性選択方式３の欄に示した属性を用いた視覚化結果を図5.3に示す．第2属性である"肥満度"は，値が離散的であるため色で表し，第3属性の"身長"をY軸に対応させて表現している．健康診断結果において，全ての属性に対して総合的に相関が高いものは体重であり，視覚化結果からは"身長に対して体重の比が大きい人は，肥満度が高い"という一般的な肥満の定義を直観的に理解することができる．

５．４　考察

１組の属性間の相関を把握するために相関係数行列は有効である．しかしながら，特にデータが持つ属性数が多い場合に，複数属性間の相関関係を把握することは難しいと考えられる．本稿で示した各方式によって，有効な相関関係を全て抽出することはできないと思われるが，相関が認められる属性群の例示ができるため，分析初期におけるデータの特徴把握に有効であると考えられる．以下に各方式についてまとめる．

属性選択方式１によって線形的な相関が最も高い（最大相関係数を有する）属性対に注目して，その属性と相関が高い属性群をユーザに提示することができる．

属性選択方式１が，最大相関係数を中心とした相関関係に着目した方式であるのに対し，属性選択方式２は，部分属性相互の相関に着目した方式である．２つ以上の部分属性間の相関をより強調して，相互に相関が高い属性群をユーザに提示できる．

属性選択方式３によって，データが持つ全属性に対して相関が高い属性およびその属性と相関が高い属性群をユーザに提示することができる．全ての属性対の相関を均等に評価する際に有効であると考えられる．

各方式で選択された属性を図形の配置，色，大きさなどに対応させて多次元視覚化表示することにより，２属性間の相関係数や通常のXYプロットでは読み取ることが難しい複数属性間の関連性を直観的に理解することができる．視覚化は，上述の特徴をもとにした複数属性間の相関の高さや特異点，分布状況などの把握に有効であると考え

られる．

エンドユーザにとって有効なパターンは，データや分析目的（どのような情報が欲しいか），背景知識などによって異なり，正解はない．特に，相関が高い属性間の関係等は既知である場合がある．したがって，各方式の適用により相関が認められる属性群を選択し，視覚化してユーザに提示することは，データに内在する有効なパターンを発見するための初期段階として有益であると考えられる．さらにユーザの背景知識などを考慮し，インタラクション等を含めた方式を検討することで，ユーザ毎に異なる要求に応えうる結果の提示ができると考えられる．例えば，ユーザが分析上注目した属性をシステムに与えて，その属性と相関が高い属性を中心に属性群を選択する方式が考えられる．

また，本研究では相関係数をデータ特徴（属性選択基準）として利用したが，意思決定者が判断する際に，より詳細な分析を行いたい場合には，相関係数の信頼度などの統計的評価が必要であると考えられる．さらに離散的な値を持つ属性も同様に扱ったが，特に名義尺度の場合には数量化理論を用いる必要がある．視覚化に際して，離散値を持つ属性を散布図的に表現した場合には図形が重複し，個々の図形の認識が難しいという問題が生じるため表現パターンについての検討も必要である．表現パターンとして，樹形図，包含図，網状図などの検討を考えている．

5　おわりに

2属性間の相関を示す相関係数行列を利用して様々な関連性を考慮して複数の属性群を選択する方式を検討し，視覚化への適用について考察した．相関係数行列から読み取ることが難しい複数属性の関連性の把握が可能であることを示した．

今後は，ユーザとのインタラクションや統計的評価値の適用，表現パターンの高度化をはじめとし，決定木や相関係数以外の手法の適用について検討を進め，高度な分析の支援を目指す．

＜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]M.Estter, H.P.Kriegel, X.Xu: "A Databese Interface for Clustering in Large Spatial Databases ", Int. Conf. on Knowledge Discovery and Data Mining, pp.94-99, 1995.

[4]T.Fukuda, Y.Morimoto, S.Morishita, T.Tokuyama : "Mining Optimized Association Rules for Numeric Attributes", Proceedings of the 15th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp.182-191, 1996.

[5]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.

[6]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

[7]K.Kurokawa, et al. : "Information Visualization Environment for Caracter-based Database Systems ", Proceedings of The First International Conference on Visual Information Systems, pp.38-47, 1996.

[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.

[9]浅野: "統計・分析手法とデータの読み方", 日刊工業新聞社, 1992.

[10]浜田，岸本: "SAS/INSIGHTによるデータ分析の教育", 統計数理 Vol.44, No.2, pp149-162, 1996.

[11]SAS Institute: "SAS/STATソフトウエア ユーザーズガイド Version 6 First Edition", SAS Institute, 1995.