1Z－4

視覚的な知識発見のための数値データの色情報による可視化の検討

吉吉健太郎　岩佐英彦　竹村治雄　横矢直和

奈良先端科学技術大学院大学

1　はじめに

近年，大規模なデータベースから有用な知識を獲得する技術としてデータマイニングが注目されている [1]．データマイニングにおいて獲得される知識の代表例として属性間の相関を表す結合ルール [2] が挙げられるが，従来の結合ルールの計算機による自動的発見の手法では，属性値が離散的な記号データに限られるものが主である．

これに対し本稿では，連続的な数値データに対する結合ルールを対象とする．数値属性間の結合ルールの発見とは，複数の数値属性において，データの分布の偏りに強い相関が存在する区間を見い出すことである．従来の結合ルール発見アルゴリズムを拡張してこの問題を解く方法も考えられるが，我々は人間の視覚的な認識能力を用いて属性間の相関を発見する手法に着目する．この手法は，色情報などの利用によって数値データを可視化し，データの偏りや属性間の関連性などを視覚的に認識可能な形に変換する．そして，人間の優れたパターン認識能力によってデータベースに内在する規則性を発見しようとするものである．本稿では，属性値に濃淡色を割り当てることでデータベースを可視化し，規則性の発見を行なう手法を提案する．また，提案システムを用いて視覚的に規則を発見する様子を示す．

2　数値結合ルール

属性値がすべて連続値をとるデータベースを考える．A, B をデータベース中の属性とし，各属性値の区間を $[a_l, a_h], [b_l, b_h]$ で表すとき，

$$(a_l < A < a_h) \Rightarrow (b_l < B < b_h)$$

の形で定義される規則を数値結合ルールと呼び，規則の左部を条件部，右部を結論部と呼ぶ．また，$[a_l, a_h]$ を条件区間，$[b_l, b_h]$ を結論区間と呼ぶ．数値結合ルールは，「属性 A の値が条件区間に含まれれば属性 B の値は結論区間に含まれる」ことを表す．

従来の結合ルールの評価尺度としてはサポートと確信度が一般的である [2]．サポートとは，ルールを満たすデータが全データに対して占める割合のことであり，確信度とは，ルールを満たすデータが条件部を満たすデータに対して占める割合のことである．これらの値が高いほど一般性の高い規則となる．連続値属性においてもこれらの評価尺度が利用可能である．

数値結合ルール発見のために数値属性の区間を設定することを考える場合には，ルールを満たすデータは狭い区間に存在することが望ましい．そこで，新たな評価尺度として，「結論区間として取り得る区間に対して結論区間が占める割合」，すなわち「結論区間の狭さ」を加える．以後，これを区間占有率と呼ぶ．区間占有率は，サポートや確信度とは逆で，値が低いほど有用な規則であることを表す．

3　視覚的知識発見手法

3.1　数値結合ルールの視覚的発見

本研究における視覚的な知識発見とは，可視化されたデータベースを人間が観察することによって，前節で述べた3つの尺度による評価が高くなるような数値結合ルールを見つけ出すことに相当する．この実現のためには，属性間の値の分布の相関が視覚的に把握できる必要がある．さらに，その相関の成り立つデータに対する3つの評価尺度を視覚的に確認できなくてはならない．以下では，これを実現するシステムを提案する．

3.2　提案システムの概略

開発した視覚的知識発見システムの概観を図1に示し，以下で各エリアについて説明する．

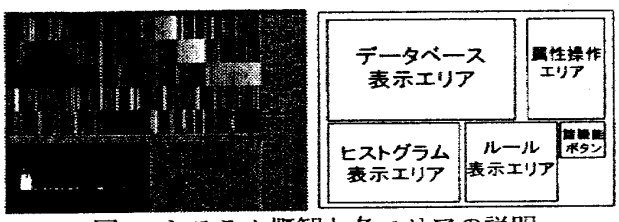


図 1: システム概観と各エリアの説明

[データベース表示エリア]

可視化したデータベースを表示する領域である．可視化の手順は以下の通りである．まず，各属性に対してその属性値を1～100の範囲で正規化する．次に正規化された属性値に対して濃淡色 (黒＝1，白＝100の100段階) を割り当てる．例を挙げると，図1のデータベース表示エリアには5つの属性をもつデータベースが表示されている．このとき，1つのデータは5つの属性に対応する濃淡色で表現される一本の縦棒として可視化されている．

A visualization method of numerical data for visual knowledge discovery

Kentaro Kichiyoshi, Hidehiko Iwasa, Haruo Takemura, and Naokazu Yokoya

Nara Institute of Science and Technology (NAIST)

8916-5 Takayama, Ikoma, Nara 630-0101, Japan

[属性操作エリア・ルール表示エリア]

属性操作エリアは, ある属性についてデータをソートしたり, ソートした属性で区間を設定することで, 対話的に条件部や結論部の決定を行なう領域である.

ルール表示エリアは, 得られた数値結合ルールとその評価値を表示する領域である.

[ヒストグラム表示エリア]

結論区間の設定を補助するために属性 (濃淡) 値ヒストグラムを表示する領域である.

有用な数値結合ルールは, 区間占有度の定義より, 結論部の指定区間が狭いことが条件となる. 「指定区間が狭い」とは, 取りうる濃淡色の範囲が狭いことに相当する. そこで結論部属性値の濃淡値ヒストグラムを用意し, より度数が高く狭い区間設定を行なえるようにする.

3.3 提案システムによる数値結合ルール発見手順

前述の評価尺度による評価が高い結合ルールを提案システムで発見するための操作は以下の通りである.

(1) 条件部・結論部となる属性の選択 属性を順に選び, 属性値によるソートを行なう. このとき, 他の属性において強い色の偏りがある場合, すなわち, 類似する色が帯状に出現している場合に, ソートされている属性を条件部属性, 色に偏りがある属性を結論部属性とする. 色の偏りがある表示部分の幅が広いほどサポートの高いルールが得られる. また, 確信度の高いルールを得るためには, その部分の色の偏りが強い方が良い. このような条件を満たす区間を, 各属性をソートしながら対話的に発見する.

(2) 条件区間の設定 上記で発見した結論部属性において色の偏りが存在する区間に対して, それに対応する条件部属性の区間を条件区間とする.

(3) 結論区間の設定 設定された条件区間に含まれるデータに対して, 結論部属性のソートを行ない, 濃淡値に偏りのある区間を結論区間として選択する. この際, 区間占有率の値を小さくするために, 濃淡値のヒストグラムを表示して補助的に利用する.

4 提案システムを用いた結合ルール発見例

提案システムを用いて, 実際に

$$(x_l < X < x_h) \Rightarrow (y_l < Y < y_h)$$

で表される数値結合ルールを発見する例を示す. 図2が初期状態のデータベース表示であり, 5つの属性を持つ800個のデータが表示されている. なお, このデータベースは, 下記の規則

(30 < 属性 2 < 60) ⇒ (1 < 属性 5 < 10)
サポート 0.3, 確信度 0.8, 区間占有率 0.1

が成り立つようにランダムに配置したものである.

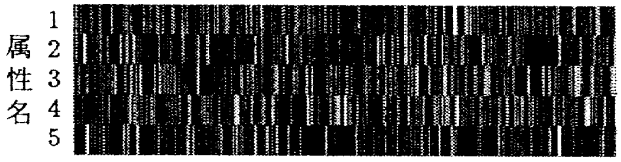
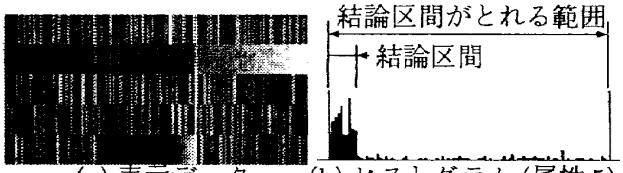


図 2: 初期状態

まず, 条件部・結論部となる属性を選択する. 図3は属性2についてソートを行なった結果であるが, 属性5において大きな色の偏りのある区間が見られる. ゆえに, 条件部属性として属性2を選択し, 結論部属性として属性5を選択する. 条件区間は, 属性5の色の偏りのある部分に対応して, 図3のように設定する.

図 3: 条件区間の設定

最後に結論区間の設定を行なう. 条件区間内のデータを結論部属性として選択した属性5についてソートする (図4(a)). その後, ヒストグラムを参考にしつつ, 結論区間を設定する (図4(b)).



(a) 表示データ　(b) ヒストグラム (属性5)

図 4: 条件区間内での属性5の選択とヒストグラム

このとき発見されたルールとその評価値は,

(30 < 属性 2 < 59) ⇒ (1 < 属性 5 < 9)
サポート 0.24, 確信度 0.74, 区間占有率 0.11

となった. すなわち, 発見例ではデータベースに内在する規則をほぼ正しく発見できたと考えられる. このことは, 提案システムによる視覚的な知識発見の可能性を示すものであると言える.

5 まとめ

可視化されたデータベースから人間の視覚能力を利用して数値結合ルールを発見する手法を提案した. 今後は, より効果的なデータベースの可視化手法を検討し, システムの高度化を図る予定である.

参考文献

[1] 河野浩之: "データベースからの知識発見の現状と動向", 人工知能学会誌, Vol.12, No.4, pp.497-504, 1997.

[2] 福田, 森本, 森下, 徳山: "データマイニングの最新動向", 情報処理, Vol.37, No.7, pp.597-603, 1996.