1N－7

圧縮情報量密度に基づく発見ルール集合の可視グラフ構造への変換

石原 寛紀　鈴木 英之進
横浜国立大学工学部電子情報工学科

1 はじめに

データベースからの知識発見 (Knowledge Discovery in Databases: KDD) は, 近年における大規模データ数の急速な増加を背景に, 活発に研究が行なわれている. KDD においてルール発見は, データ集合中に存在する関係の部分集合をルール形式で発見することである. ルールは,"if A then B" 形式, つまり, ある条件 A が成立するときに, 条件 B が高い確率で成立する $A \rightarrow B$ という規則をいう.

一般的に, ルールはデータ集合から極めて多数発見されるので, ルールの興味深さをより詳細に評価するために, ルールの個数を減らし, 人間にとって理解が容易な知識形式で表現することが望ましい. 従来のルール発見手法 [Smyth 92] や, 得られた知識を視覚的に出力する手法 [Jensen 96] は, 出力表現がこの点を満たしてない欠点があった.

本稿では, 与えられた発見ルール集合を, 単純な構造でデータの情報を多く圧縮する可視グラフ構造に変換する手法を提案する. 農業統計データを用いた実験の結果, 本手法の有効性が示された.

2 対象問題

本稿で扱うデータは, 複数個の属性で記述される例の集合で構成される. ただし, 連続値属性は, 前もって離散値属性に変換されているとする. また, ある属性が一つの属性値をとる事象を, アトムと呼ぶことにする.

ルール $r(\mu)$ は, μ 個の連言アトムで表される前提部 Y_μ が成立する時に, 単一のアトムで表される結論部 $X = x$ がある確率で成立するプロダクションルールを表す. $r(\mu)$ は以下のように定義される.

$$r(\mu) \equiv Y_\mu \rightarrow X = x \quad (1)$$

$$Y_\mu \equiv (Y_1 = y_1) \wedge (Y_2 = y_2) \wedge \cdots \wedge (Y_\mu = y_\mu) \quad (2)$$

ルール集合 R を表すグラフ $G(R)$ は, 前提部 Y_μ において出現頻度が高いアトムを高い順に同一アトムにまとめ, 各前提部から結論部 $X = x$ に向かう方向にアークを引いて表現される. 本手法では, まず入力されたルール集合 R をグラフ $G(R)$ に変換する. 次に,$G(R)$ から, 簡単な構造で, データの情報を圧縮する可視グラフ構造 $G(R')$ を出力する.

[0]Transforming Discovered Rule Set into Visual Graph Structure based on Compressed Entropy Density
Hiroki Ishihara and Einoshin Suzuki
Division of Electrical and Computer Engineering, Faculty of Engineering, Yokohama National University
79-5 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan

3 可視グラフへの変換

3.1 グラフの評価規準

グラフ $G(R)$ を評価する規準値を提案する. 圧縮情報量密度 $D(G(R))$ は, グラフ $G(R)$ の持つ情報量 $I(G(R))$ とグラフを表すのに必要な記述長 $L(G(R))$ の商である.

$$D(G(R)) \equiv \frac{I(G(R))}{L(G(R))} \quad (3)$$

以下, $I(G(R))$, $L(G(R))$ について述べる.

(1) 式のルール $r(\mu)$ について, 前提部および結論部のデータ集合全体に対する存在確率をそれぞれ $p(y)$, $p(x)$ とする. この時, $r(\mu)$ が持つ情報量 $J(r(\mu))$[Smyth 92] は, 以下の式で与えられる.

$$J(r(\mu)) \equiv p(y)\left(p(x|y)\log_2\frac{p(x|y)}{p(x)} + p(\bar{x}|y)\log_2\frac{p(\bar{x}|y)}{p(\bar{x})}\right) \quad (4)$$

本稿では, グラフの持つ情報量 $I(G(R))$ を以下の式で定義する.

$$I(G(R)) \equiv \sum_{r_i(\mu_i)\in G(R)} J(r_i(\mu_i)) \quad (5)$$

グラフを表すのに必要な記述長は, グラフ中に, どのアトムが連言をいくつとるかという情報に基づき計算される. ただし, 本手法では, 結論部およびアークの記述長は無視する. よって, A をデータ中に含まれるアトムの種類数とすると, グラフ $G(R)$ における記述長 $L(G(R))$ は, 以下の式で定義される.

$$L(G(R)) \equiv \sum_{r_i(\mu_i)\in G(R)} \mu_i \log_2 A \quad (6)$$

3.2 探索手法

与えられた発見ルール集合をグラフ構造に変換し, 圧縮情報量密度に基づいてグラフの持つ情報を圧縮し, 可視グラフ構造に変換する探索手法を提案する.

入力するルール集合を R, 探索時のルール集合を R_c, R_c から x 番目のルールを削除したルール集合を $R_{c,x}$ とする. 探索は, 与えられたルール集合グラフ $G(R_c)$ から x 番目のルールを削除したグラフ $G(R_{c,x})$ の中で最大の圧縮情報量密度を持つグラフ $G(R_m)$ を返す山登り法を採用している. なお, 最終的に出力するルール数 θ は, あらかじめ設定しておく. またルール集合 R_c に含まれるルール数を $|R_c|$ で表す.

1. (Set) $R_c \leftarrow R$, $D(G(R_m)) \leftarrow 0.0$
2. (ルール集合を圧縮)
   (a) while($|R_c| > \theta$)
   (b) for each x from 1 to $|R_c|$
   (c) if $D(G(R_{c,x})) > D(G(R_m))$, $R_m \leftarrow R_{c,x}$
   (d) $R_c \leftarrow R_m$
3. (Return) Return $G(R_c)$

4 農業データへの適用実験

本手法を平成 4 年度農業市町村別基礎統計データに適用した. 使用したデータは, 平地の農業地域に対して, 森林や傾斜地が多く, 平地が少ない市町村である全国中山間地指定地域 1748 例に関するものである. 各例に対して属性は 148 個存在する. ただし, 連続値属性は, 欠落値と "0" をそれぞれ一つの属性値とし, それら以外の値を等頻度離散化方式により { 低, 中, 高 } に離散化した. このデータからルール発見手法 [鈴木 99] によって出力された 450 個のルール集合を入力とし, ルール数 15 個の可視グラフを出力させた. 図 1に「一戸当り生産農業所得が高い」ルールの可視グラフを, 図 2に「専従一人当り生産農業所得が高い」ルールの可視グラフを示す. 図 1, 2を見ると, 出力された可視グラフは, 各ルールが妥当な意味を持ち, 単純で見やすいことがわかる. また, 本結果を複数の農業の専門家に見せ, 同じ評価を得ている.

5 おわりに

本稿では, 多数発見されるルール集合を, データの持つ情報を単純な構造で圧縮し, 可視グラフに変換する手法を述べた. さらに本手法を, 農業に関する 1748 市町村の統計データに適用した. 得られた可視グラフ構造は, 農業の専門家にとっても妥当で見やすいものであり, 本手法の有効性が示された.

参考文献

[Jensen 96] Jensen, F. V.: An Introduction to Bayesian Networks, *Springer-Verlag(1996).*

[Smyth 92] Smyth, P. and Goodman, R. M.: An Information Theoretic Approach to Rule Induction from Databases, *IEEE Trans. on Knowledge and Data Eng.,* vol.4, No.4, pp.301-316(1992).

[鈴木 99] 鈴木 英之進: データベースからの特徴的ルール発見のための一般性と正確性の信頼性同時評価方法, 人工知能学会誌, Vol. 14, No.1, pp. 139-147 (1999).

謝辞

本研究の一部は、農林水産省一般別枠研究「増殖情報ベースによる生産支援システム開発のための基盤研究」による.

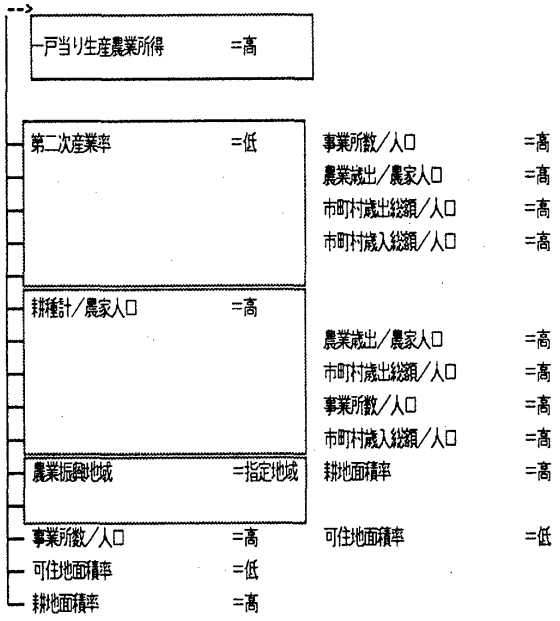

図 1: $\theta = 15$ としたときの「一戸当り生産農業所得=高い」結論部を持つ可視グラフ

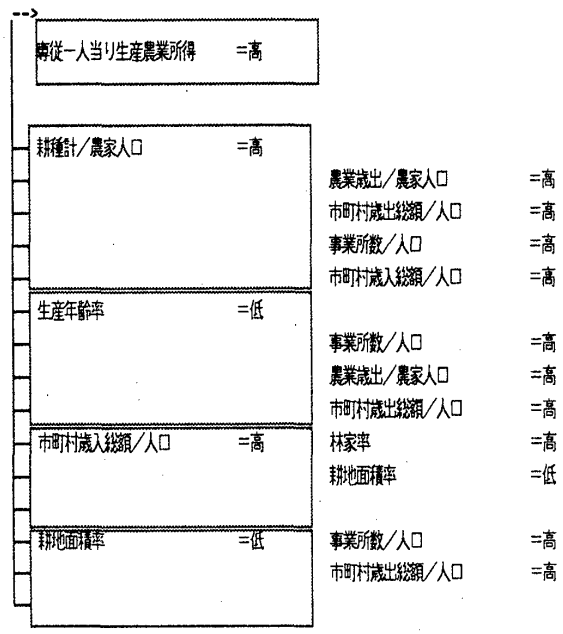

図 2: $\theta = 15$ としたときの「専従一人当り生産農業所得=高い」結論部を持つ可視グラフ