2002年日本オペレーションズ・リサーチ学会
秋季研究発表会

データの論理的解析における ルール集合の生成について

(申請中)　京都大学　*辻 弘貴　TSUJI Hiroki
01001374　京都大学　茨木 俊秀　IBARAKI Toshihide

1　はじめに

近年, 大量のデータが簡単にしかも安価に蓄えられるようになり, そこから意味のある知識を抽出するための科学的手法の重要性が増している. 本研究ではルール形式の知識の抽出を試みる.

本研究で扱う個々のデータ $\boldsymbol{a}=(a_1,a_2,...,a_n)$ は n 個の属性に対して値を持つ n 次元ベクトルとして表現される. 各 $\boldsymbol{a}$ は正もしくは負いずれかの決定クラスに属し, 正データ集合を $\boldsymbol{P}$, 負データ集合を $\boldsymbol{N}$ とすると $\boldsymbol{P}\cup\boldsymbol{N}=\boldsymbol{S}$, $\boldsymbol{P}\cap\boldsymbol{N}=\emptyset$ が成り立つ. 各 $\boldsymbol{a}$ のクラスを $\Omega(\boldsymbol{a})$ と書き, 1 (正クラス) あるいは 0 (負クラス) の値をとる [2, 3].

本研究で扱うルール $r=(\pi,\omega)$ は条件部 π と結論部 ω から成る. 条件部は特徴と呼ばれるものの連言表現からなり, 結論部は $\omega\in\{0,1\}$ である. ルールは「条件部が成り立つようなデータは ω のクラスに属する」ことを表現し, 支持度 μ と信頼度 ν によって確率的に評価される [1].

本研究ではルール単独の評価よりも, ルールの集合として評価することを考え, データの正負の分類に有効な生成法を提案する. 生成されたルール集合の評価には, その集合から分類器である決定木を構成し, 未知のデータに対する誤分類率を用いる.

計算実験においては, あらかじめルール集合を定めることなく決定木を構成する通常の方法と誤分類率を比較し, 提案手法の有効性を確かめた.

2　用語の定義

データ集合 $\boldsymbol{S}$ の持つ n 個の属性は, 数値属性もしくは記号属性のいずれかに分類される. 数値属性は大小関係を持つ数値で表現され (例：身長, 体温, 検査値), 記号属性は大小関係を持たない記号で表現される (例：色, 匂い, 味).

個々の特徴は, それぞれ 1 つの属性に対して定義される. 数値属性に対しては閾値特徴 $x=T(i,c)$ ($a_i\geq c$ かどうかを判定) , およびインターバル特徴 $x=I(i,c,c')$ ($c\leq a_i\leq c'$ かどうかを判定) と呼ばれる 2 種類が定義され, 記号属性に対しては部分集合特徴 $x=S(i,A)$ ($a_i\in A$ かどうかを判定) が定義される. 属性 i の特徴 x はそれぞれの性質に従いデータ $\boldsymbol{a}$ の i 番目の属性値 a_i に対して $\{0,1,*\}$ のいずれかを出力する. すなわち, $x[\boldsymbol{a}]\in\{0,1,*\}$ であって, $x[\boldsymbol{a}]=1(0)$ は特徴 x が成立する (成立しない) ことを示し, $x[\boldsymbol{a}]=*$ はどちらとも言えないことを意味する. ここで, $\rho(x)=|\{(\boldsymbol{a},\boldsymbol{b})\mid(x[\boldsymbol{a}],x[\boldsymbol{b}])\in\{(0,1),(1,0)\},\boldsymbol{a}\in P,\boldsymbol{b}\in N\}|$ と定義し, $\rho(x)$ を特徴 x の分離可能ペア数と呼ぶことにする.

各特徴 x はブール変数とみなすことができるが, その否定を $\bar{x}$ と記す. x, $\bar{x}$ を合わせてリテラルと呼ぶ. ルール r の条件部 π はリテラルの連言表現から成り, $\pi=w_1w_2\ldots w_k$ のように書かれる (ここで, 各 w_i は一つのリテラル x あるいは $\bar{x}$ である). $\pi[\boldsymbol{a}]=w_1[\boldsymbol{a}]w_2[\boldsymbol{a}]\ldots w_k[\boldsymbol{a}]$ は 0, 1 あるいは$*$の値をとる. $\pi[\boldsymbol{a}]=1$ のとき,「$\boldsymbol{a}$は rの条件部を満たす」という. r の結論部 ω は 0 あるいは 1 の値をとる. $\boldsymbol{a}$ が r の条件部を満たし, かつ $\Omega(\boldsymbol{a})=\omega$ であるとき,「$\boldsymbol{a}$はルール r を満たす」という [2, 3].

ルールの性能を評価するため, ルール $r=(\pi,\omega)$ に対し, 支持度 $\mu(r)$ および信頼度 $\nu(r)$ を以下のように定義する. ただし, ルール r の条件部を満たすデータ集合を $\boldsymbol{S}_\pi$ とする [1].

$$\mu(r)=\frac{|\boldsymbol{S}_\pi|}{|\boldsymbol{S}|},\quad \nu(r)=\frac{|\{\boldsymbol{a}\in\boldsymbol{S}_\pi\mid\Omega(\boldsymbol{a})=\omega\}|}{|\boldsymbol{S}_\pi|}.$$

3　提案するルール集合生成法

今回提案するルール集合生成アルゴリズムは, 出力 $\omega=0$, 1 のそれぞれに対して, 条件部のリテラルの数が $1,2,\cdots,d_{\max}$ ($d_{\max}$ は定数) のルールのうち, μ, ν が比較的高い, 有効なルールの集合を欲張り法的な手法によって求めるものである.

アルゴリズムの基本方針は, 条件部のリテラル数が d であるルール集合 R_d を, R_{d-1} のルールそれぞれの条件部に新たにリテラルを連言に付け加えることによって求める. しかし, 全てのリテラルを付け加えるのではなく, R_{d-1} のルールそれぞれに対して, 分離可能ペア数の高い特徴を K_d (定数) 個選び出し, リテラルの形で付け加える. つまり, $|R_d|=\prod_{i=1}^{d}K_i$ となる. 最後に $r\in\bigcup_{d=1}^{d_{\max}}R_d$ のうち, 支持度, 信頼度が, $\mu_{\min}$, $\nu_{\min}$ (入力パラメタ) 以上であるルールのみを, R として出力する.

このアルゴリズムにより「$a_6\geq 2.5$ かつ $a_2\geq 3.5$ なら, 正データ (支持度 30%, 信頼度 97%)」というようなルールの集合を見つけることができる.

4 ルール集合の評価

生成されたルール集合の分類能力を評価するために, 得られたルール集合から分類器を構成し, その性能を測る. 分類器は, 既知のデータから抽出された知識に基づいて構成され, クラスが未知のデータに対して, それを予測するものである. 分類器は誤分類率によってその性能を評価する.

本研究では分類器として決定木を用いる [4]. 決定木とは, 条件分岐によりデータの存在する領域を次々に分割し, 任意のデータを 1 つの決定クラスに分類するものである. 決定木の例を図 1 に示す. 今回, 決定木構成アルゴリズムには C4.5 を採用した [4].

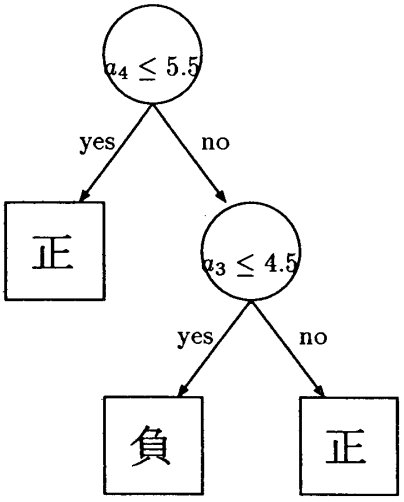

図 1: 決定木の例

従来の決定木は 1 つの特徴を節点とするが, ルール集合の評価には, その集合に含まれるルールの条件部, つまり, リテラルの連言表現を節点とした決定木を構成する. この決定木の誤分類率を測定することにより生成されたルール集合の評価を行う. すなわち, 生成されたルール集合がデータの特性をうまく記述するものであれば, ルールの条件部は, 正データと負データの分類のための本質的な情報を担っており, それらを用いて決定木を構成すれば低い誤分類率を達成できると考えられる.

5 実験結果

実験方法は, まずデータ集合をトレーニングデータとテストデータにランダムに 1 対 1 に分ける. 次に, トレーニングデータを直接 C4.5 に入力して, 節点に特徴を配置した従来型の決定木と, トレーニングデータから生成したルール集合の条件部を配置した決定木の 2 種を構成する. 次に, それぞれの決定木にテストデータを入力し, 誤分類率を測定, 比較する. 実験には, ベンチマークとして広く用いられているデータ集合をいくつか用いたが, そ

表 1: いくつかのデータ集合に対する誤分類率

決定木	BCW	AUS	IONO	CAR
従来法	4.62%	16.76%	10.85%	6.83%
提案法	3.65%	14.84%	11.07%	2.52%

の中から,

BCW 乳癌の診断データ (データ数 683)
AUS クレジットカードの査定データ (データ数 383)
IONO 電離層の反射に関するデータ (データ数 351)
CAR 車の評価に関するデータ (データ数 1728)

の 4 つのデータの実験結果を表 1 に示す. 表の値は, トレーニングデータとテストデータを分ける際, 異なるランダムな分け方をした 10 回の実験の誤分類率の平均値である. 結果として, 一部のデータでやや劣るものが見られるものの, 概ね提案手法がやや良い結果を出している. IONO のデータで悪い結果となっているのは, データ自体に雑音が多いため, 過学習していると考えられる.

6 おわりに

ここでは紙面の都合上, ルール集合の生成法として 1 つしか述べなかったが, これ以外にもいくつかのルール集合生成法を考案している. それらから構成された決定木の誤分類率も求めている. また, 提案手法で生成されるルール集合は大きくなる傾向が強いので, C4.5 に入力する前に, 有効だと思われるルールを色々な手法によって選別し, ルールの数を減らした上で (事前処理)C4.5 に入力するという方法も試みた. その結果によると, 事前処理を加えた方法が良い結果を残しており, 冗長なルールが効率よく省かれていることを示している.

参考文献

[1] R. Agrawal, T. Imielinski and A. Swami, Mining association rules between sets of items in large databases, *International Conference on Management of Data (SIGMOD 93)*, (1993) 207–216.

[2] 原口 和也, 茨木 俊秀, データからの知識獲得における常識ルールと例外ルールについて, 2001 年度 日本オペレーションズ・リサーチ学会秋季研究発表会アブストラクト集, (2001) 214–215.

[3] Mii, S.: *Feature determination algorithms in the analysis of data*, 京都大学修士論文, 2001.

[4] J.R.Quinlan, *C4.5: Programs for machine learning*, Morgan Kaufmann, 1993.