散布図内の小領域を特徴づける属性組み合わせ発見手法の提案

佐々木郁哉[1]　嶋田香[2]　荒平高章[1]

概要：本論文では，相関ルール「P→Q (If P then Q)」においてQが注目しているX，Yの2変数の特徴的な分布を与える小領域となる拡張したルール発見手法について提案する．提案手法は，２変数X，Yの散布図内の小領域を特徴づける属性組み合わせPを発見する手法の提案となる．この手法に関して，SSDSE（教育用標準データセット）を用いた評価実験を行った．その結果，2変数X，Yによる散布図内における小領域を特徴づける属性組み合わせを発見することができた．今後は，この発見した属性組み合わせにおけるデータを詳細に検討し，新たな知識の創出を目指す．

キーワード：散布図，属性組み合わせ，進化型計算

1. はじめに

相関ルールは「P→Q (If P then Q)」の形式で表現され，Qがクラス属性のときは，ルールベースのクラス分類に用いることができる．これまでにクラス分類のためのルール発見法とクラス分類法に関する多くの提案がなされており，大きな成果をあげている．相関ルールを用いたクラス分類では，ルールのもつ表現形式から可読性があり，クラス分類をどのような理由で行っているかの解釈に役立つと考えられる．最近，相関ルールを用いた連続値の予測への関心が高まってきているが，手法に関する報告は多いとはいえない．

著者らはQが注目している連続1変数Xにおける特徴的な小分布となる場合のルールを定義して発見する手法を成果蓄積型の進化型計算手法を用いて提案している[1]．小分布を散布度が小さく鋭い局所的な分布とすると，学習データから発見されたルール集合により，テストデータの連続変数値をルールベースで予測することが考えられる[2]．本論文では，Qが注目しているX，Yの2変数へ拡張したルール発見手法について提案する．提案手法は，２変数X，Yの散布図内の小領域を特徴づける属性組み合わせPを発見する手法の提案となる．本手法を用いることで，1変数によるルール発見では発見できないような新たな属性組み合わせを発見することを可能とし，社会科学的な統計データや医療統計データといったさまざまなデータを用いて新たな知見を創出することを目的とする．

2. ルール発見法

2.1 ルールの表現と指標

多数の属性に関する値とこれらに関連付けられた連続変数を有するTable1の形態のデータベースから，属性間に見られる特定組み合わせの時に観察される連続変数の特徴的な小分布を IF～Then～型のルール形式で発見する．A_i, (i=1, ..., n)は1，0の２値をとり、欠損値はないものとする．また，X，Yは連続値をとる変数とする．本論文で扱うルールを以下のように定義する．

［Associative local distribution rule］
$(A_j=1)\wedge\cdots\wedge(A_k=1)\rightarrow(m_X(r),\ s_X(r),\ m_Y(r),\ s_Y(r))$
ここで，$m_X(r)$と $s_X(r)$，$m_Y(r)$と $s_Y(r)$は，ルールrの前件部を満たすレコードのX，Yの平均値と標準偏差である．ルールを簡単に $P\rightarrow(m_X,\ s_X,\ m_Y,\ s_Y)$と表す．ここで，$P=(A_j=1)\wedge\cdots\wedge(A_k=1)$であり，Pを $A_j\wedge\cdots\wedge A_k$と略記する．

Table 1 : Example of database.

ID	A_1	A_2	A_3	A_4	X	Y
1	1	0	1	1	1.03	1.44
2	1	1	1	0	0.85	1.96
3	0	1	1	1	-1.89	-0.13
4	0	0	1	1	1.79	-1.09
5	0	1	1	1	0.10	-0.12
6	1	1	0	1	0.14	-0.08
7	0	1	1	0	-0.49	-0.14
8	1	0	1	0	1.19	-0.19

ルールrに対して，その評価に利用できるレコードの総数をN(r)，そのうち前件部を満たすレコード数を $N_{x,y}(r)$とし，

$$\text{Support(r)}=\frac{N_{x,y}(r)}{N(r)} \quad (1)$$

をルールの出現頻度の指標とする．

1 九州情報大学
Kyushu Institute of Information Sciences
2 福岡看護大学
Fukuoka Nursing College

また, 前件部を満たすレコードの X, Y の値の平均を $m_X(r)$, $m_Y(r)$とし，標準偏差を $s_X(r)$，$s_Y(r)$とする．
本論文では，r が以下の条件を満たすとき，興味深いルールと定義する．

$$Support(r) \geq supmin \quad (2)$$

$$s_X(r) \leq smax \quad (3)$$

$$s_Y(r) \leq smay \quad (4)$$

ここで，supmin，smax，smay は予めユーザが定めておく値である．

2.2 ルールの発見方法

ルールを発見するためには，i)ルールの前件部の設定，ii)指標の算出，iii)興味深いルールかどうかの判定，の 3 段階を経る必要がある．遺伝的ネットワークプログラミング(GNP)を応用したルール発見法[1], [2]では，i), ii)を GNP を応用したノードの接続と遷移により扱い，iii)の判定に基づいて GNP 個体の適合度を算出する．一般に，GNP を含めて進化論的計算手法では個体を解として問題を扱い，進化による世代の経過により最適な適合度の個体を求めて解とする．一方，提案手法ではルールが解となるが，これを個体として扱うのではなく，各世代に存在した個体集合から世代継続的に未抽出のルールを解として抽出し続ける成果蓄積方式となっている．提出方法は，文献[1], [2]の GNP を用いた連続 1 変数 X における特徴的な小分布発見法の拡張手法として位置づけることができる．

2.3 GNP とルールの対応

ルールを構成する 1 つの属性を GNP の 1 つの判定ノード(Judgement node)で表現する．判定ノードでは指定した属性値であるかどうかを判定して，Continue または Skip に分岐していく．具体的には，属性値が 1 のとき Continue に，0 のときに Skip に接続される．また, 開始点ノード(Pointing node)は固有の順番(名前)を有しており，接続先として判定ノードを 1 つ選ぶことで, ルールの読取開始点を表す. Fig.2 は提案手法の基本的なアイデアを示したものであり，判定ノードの連結が連続変数の分布に対するフィルターの役割を果たす様子を描いている. Fig.2 の $A_1=1$ などは判定ノードの判定内容を表している．開始点ノード P_1がルールの読取開始点であり, 例えば，これと $A_1=1$,$A_2=1$ および $A_3=1$ の 4 つのノードの連結が $A_1\text{^}A_2\text{^}A_3 \rightarrow (m_{X3}, s_{X3}, m_{Y3}, s_{Y3})$ と対応する．

判定ノードの Continue 側の接続先は判定ノードと対応している．また判定ノードの Skip 側の接続先は，次の順番の開始点ノードとする．Skip 側に進む場合は，次の順番の開始ノードに遷移し，別の判定ノード群の探索を行う．判定ノードを Continue 側に連続して遷移する場合は，開始点ノードから遷移したノード数が予め指定した個数，すなわちルール前件部を構成する属性数の最大値に達したら，判定ノードの判定結果によらず次の順番の開始ノードに遷移する．Fig.2 はルール抽出を目的とした GNP 個体の基本構造を示しているが，GNP の 1 個体について，判定ノード 100 個，開始点ノード 10 個の設定が標準的である．

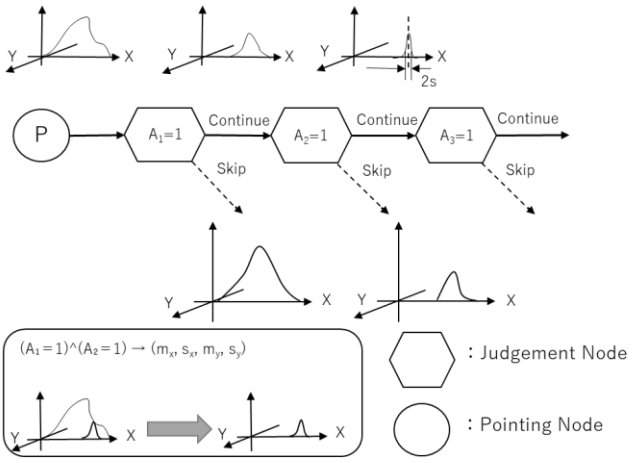

Fig.2 : Basic idea of associative local distribution rule mining.

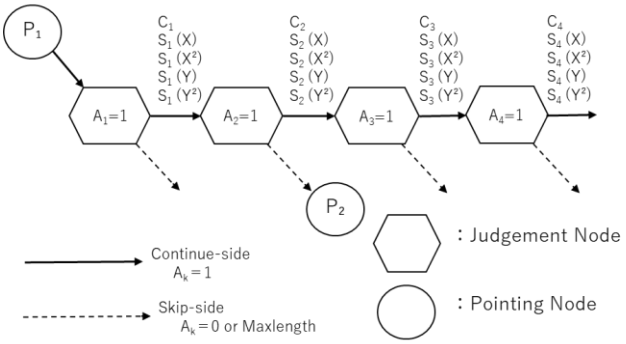

Fig.3 : Example of node connection in a GNP individual.

2.4 ルール指標の算出

データベースのレコード 1 つずつについて，GNP のノードを開始点ノードから順に遷移していく．例として，Table1 の ID＝4 は，$A_1=1$ を満たすが，$A_2=0$ なので Fig3 において P_1から遷移を始め，P_2に達する．GNP の探索は，各レコードについて第 1 番目の開始点ノード P_1からスタートし，すべての開始点ノードを始点とする遷移を終えたらレコードを更新し，また P_1に戻る．すべてのレコードを調査した後，Fig.3 の開始点ノード P_1を通過した N 個のレコードのうち，判定ノード群を Continue 側に分岐したレコード数を図のうちのそれぞれの判定ノードの位置で調査する．これらは，初期値を 0 としておいて，ノード遷移のたびに更新することで算出される．それらのレコード数のうち，属性値が開始点ノードから連続してすべて 1 であるレコード数を C_1，C_2，C_3，C_4…として求める．また，上述の C 値を更新していく過程で，属性値が開始点ノードから連続してすべて 1 であるレコードの X，Y の値の和(S(X)，S(Y))と X^2，Y^2の値の和(S(X^2)，S(Y^2))をそれぞれ判定ノードの位

置で求めるようにする．このようにして，Table2 に示すように各ルールの support，m_X，m_Y，s_X^2および s_Y^2を算出する．

Table 2 : Measurements of associative local distribution rules.

Associative local distribution rules	*support*	*m*	s^2
$A_1 \rightarrow (m_{X1}, m_{Y1}, s_{X1}, s_{Y1})$	$\frac{C_1}{N}$	$m_{X1}=\frac{S_1(X)}{C_1}$ $m_{Y1}=\frac{S_1(Y)}{C_1}$	$s_{X1}^2=\frac{S_1(X^2)}{C_1}-m_{X1}^2$ $s_{Y1}^2=\frac{S_1(Y^2)}{C_1}-m_{Y1}^2$
$A_1 \wedge A_2 \rightarrow (m_{X2}, m_{Y2}, s_{X2}, s_{Y2})$	$\frac{C_2}{N}$	$m_{X2}=\frac{S_2(X)}{C_2}$ $m_{Y2}=\frac{S_2(Y)}{C_2}$	$s_{X2}^2=\frac{S_2(X^2)}{C_2}-m_{X2}^2$ $s_{Y2}^2=\frac{S_2(Y^2)}{C_2}-m_{Y2}^2$
$A_1 \wedge A_2 \wedge A_3 \rightarrow (m_{X3}, m_{Y3}, s_{X3}, s_{Y3})$	$\frac{C_3}{N}$	$m_{X3}=\frac{S_3(X)}{C_3}$ $m_{Y3}=\frac{S_3(Y)}{C_3}$	$s_{X3}^2=\frac{S_3(X^2)}{C_3}-m_{X3}^2$ $s_{Y3}^2=\frac{S_3(Y^2)}{C_3}-m_{Y3}^2$
$A_1 \wedge A_2 \wedge A_3 \wedge A_4 \rightarrow (m_{X4}, m_{Y4}, s_{X4}, s_{Y4})$	$\frac{C_4}{N}$	$m_{X4}=\frac{S_4(X)}{C_4}$ $m_{Y4}=\frac{S_4(Y)}{C_4}$	$s_{X4}^2=\frac{S_4(X^2)}{C_4}-m_{X4}^2$ $s_{Y4}^2=\frac{S_4(Y^2)}{C_4}-m_{Y4}^2$

次いで，各開始点ノードを始点とするルールの指標を計算して，興味深さの条件を満たすかどうかを判別する．ユーザが予め与えた興味深さの条件を満たす興味深いルールが抽出された時点で，未抽出であるかどうか判断し，新規の場合，世代間に共通なルールライブラリに加える．

2.5 GNP の進化

進化時の遺伝的操作によるノード接続先変更や判定ノードを扱う属性変更は，前世代の個体内に表現されていた興味深いルールの属性組み合わせの一部を組み替えることに対応する．例えば，$A_1 \wedge A_2 \wedge A_3 \rightarrow (m_{X3}, s_{X3}, m_{Y3}, s_{Y3})$が興味深いルールの場合，$A_1 \wedge A_2 \wedge A_p \rightarrow (m_X, s_X, m_Y, s_Y)$は興味深いルールの候補として期待できる．こうした未発見の興味深いルール候補を表現可能な個体を，新しいルールを発見できた個体群に対する突然変異や交叉により得ることで次世代を構成していく．各個体の適合度(評価関数)は，未発見の興味深いルールを次世代において発見することが期待できる個体ほど大きな値をもつように設定する．具体的には，GNP 個体の内部に表現されていた興味深いルール全体の評価値の合計を用いる．各ルールの評価を，ルール指標，ルールを構成する属性数，ルールの新規性の 3 つの観点で数値化する．適合度は用いるデータの特性を考慮して設定できるが，3．評価実験では X, Y を標準化した後の値で実験を実施しており，次式を用いた．

$$F=\sum_{r \in Rd}\{a \times support(r)+\frac{b}{sX(r)+c}+\frac{b}{sY(r)+c}+n(r)+Cnew(r)\} \quad (5)$$

ここで，Rd はその個体から抽出された Eq.(2)，Eq(3)，Eq(4)を満たすルールの集合，n(r)は r の前件部を構成する属性の個数，a, b, c は定数，Cnew(r)は次式で定める定数である．

$$Cnew(r)=\begin{cases} Cnew & (rule\ r\ is\ new) \\ 0 & (otherwise) \end{cases} \quad (6)$$

なお，定数は実験的に定めている．遺伝的操作の詳細は文献[1],[2]と同様である．

3. 評価実験

3.1 評価に用いたデータ

評価実験は，SSDSE(教育用標準データセット)[3]を用いて行った．SSDSE はデータ分析のための汎用素材として，統計センターが提供しているデータセットであり，縦に地域, 横にデータ項目を並べた 2 次元の表形式データである．本実験で使用したデータの概要は以下の通りである．

・レコード数 103(福岡―鹿児島の市町村区)，属性数 70(データ項目)

・予測対象属性(X，Y)は X：非水洗化人口，Y：死亡数

本実験において，データは福岡―鹿児島の二県の市町村区(計 103 個)と男女別や年齢別，職業別などの人口割合(計 70 個)を扱った．X，Y それぞれの項目については，X を非水洗化人口，Y を死亡数として扱う．男女別や年齢別，職業別などの人口割合を 1 と 0 に 2 値化して $A_1,\ldots,A_{70}$ とし，X，Y における標準化を行った．2 値化については，人口割合の 30%<ならば 1，30%>ならば 0 というようにした．

3.2 実験の設定

発見するルールの興味深さ指標として，Eq.(2)で supmin＝0.04，Eqs.(3), (4)で smax＝0.4, smay＝0.4 とし，さらに nx(r)≦6 とした．GNP の個体数は 120 とし，1 個体あたりの判定ノードと開始点ノードの個数はそれぞれ 100，10 とした. 100 世代を 1 ラウンドとしてルール発見の 1 単位とし，1 ラウンドにおける最大ルールの蓄積数は 2000 とした．各ケースで 200 ラウンドのルール発見を行って，ルールの重複を確認し，ルール集合を得ることとした．適合度では，Eq.(5)で a＝10，b＝2，c=0.1 とし, Eq(6)で Cnew＝20 とした．プログラムは C 言語で作成し，Intel(R) Core(TM)i3-7130U CPU with 64GB RAM の計算機を用いた．

3.3 結果

実験の結果として発見されたルールの具体例として 4 つの散布図を挙げる. Fig5,6,7,8 はそれぞれが（A_{12}, A_{59}, A_{64}），（A_{10}, A_{59}, A_{64}），（A_{53}, A_{54}, A_{65}），（A_3, A_{24}, A_{54}）が小領域を特徴づける属性組み合わせとして発見された例である．

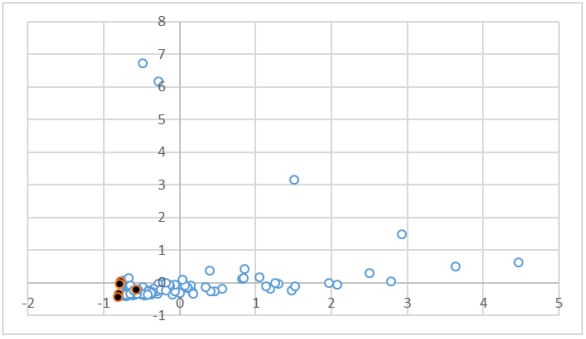

Fig.5 Example of associative local distribution rule (A_{12}, A_{59}, A_{64}).

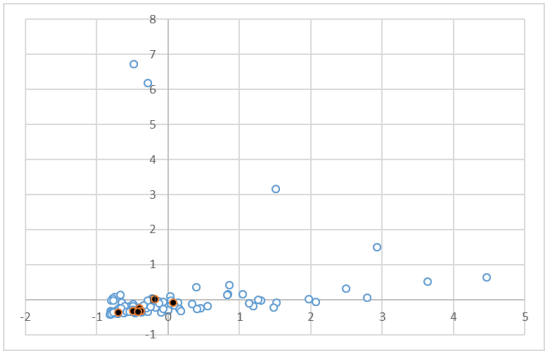

Fig.6 Example of associative local distribution rule (A_{10}, A_{59}, A_{64}).

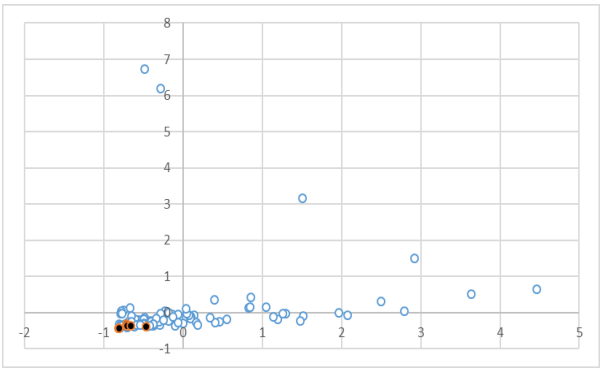

Fig.7 Example of associative local distribution rule (A_{53}, A_{54}, A_{65}).

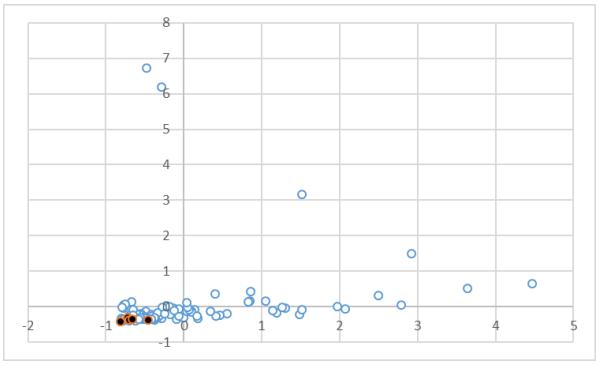

Fig.8 Example of associative local distribution rule (A_3, A_{24}, A_{54}).

Fig.5，6，7，8 いずれも●が発見された小分布である．これらの属性は具体的に，(A_{12}, A_{59}, A_{64}) = (65 歳以上人口/総人口，非労働力人口，第 3 次産業就業者数)，(A_{10}, A_{59}, A_{64}) = (15～64 歳人口（男）/総人口，非労働力人口，第 3 次産業就業者数)，(A_{53}, A_{54}, A_{65}) = (就業者数，就業者数（男），総人口（非水洗化人口＋水洗化人口))，(A_3, A_{24}, A_{54}) = (日本人人口/総人口，従業者数，就業者数（男)) である．散布図を見てもわかる通り，発見された小分布はどれも密集地帯にある．さらに，Fig.7，8 ではまったく同じ小分布が発見されている．属性(人口割合)は二つとも違うものであるため，恐らくそれぞれ異なる 3 つの属性(人口割合)全てを満たす市町村区が全く同じであったためにこのような結果が出たと考えられる．これについては人口割合を 2 値化する際の割合の線引きについて詳細に検討することで結果が変化することが考えられる．以上より，予測対象属性 (X, Y) を X : 非水洗化人口，Y : 死亡数とした場合において，様々な小領域の属性組み合わせが発見されることが明らかとなった．これらの発見された属性組み合わせに対しては，その妥当性等について今後詳細に検討していく必要がある．

4. まとめ

相関ルールは P→Q と表現されるが，Q が X，Y が注目している連続変数における特徴的な小領域となる場合のルールを定義し，データベースから発見する手法を成果蓄積型の進化計算手法を用いて提案した．評価実験は，SSDSE（教育用標準データセット）を用いて，実際に取り込んだデータから小領域を発見することに成功した．今後は，本手法によって発見した小領域について，その属性組み合わせの妥当性・有用性について詳細な検討を実施していく．

参考文献

[1] Shimada, K., and Hanioka, T., An evolutionary method for associative local distribution rule mining. In Industrial Conference on Data Mining. Springer, Berlin, Heidelberg, 2013, pp. 239-253.
[2] 嶋田香，荒平高章，埴岡隆，不完全データに対応したルールベースの連続値予測法と人工的欠損値の利用，第 42 回知能システムシンポジウム 講演論文集 E-07，2015.
[3] 統計データ分析コンペティション SSDSE https://www.nstac.go.jp/SSDSE/