知能と情報（日本知能情報ファジィ学会誌）Vol.16, No.1, pp.60-69(2004)

原著論文

エージェントによるデータベースからのルール抽出法†

領家 美奈*1　中森 義輝*2

本論文では、混合データベースからのルールを発見するエージェントについて提案を行なう。データベースには、複数の構造がしばしば混在しており、さらに、記号を持つ属性と数値を持つ属性の両方が存在する。エージェントは類似度と同定するいくつかの集合に基づいて行動しルールを同定する。得られるルールの結論部はエージェントが決定し、区間や線形モデルなどの形態をとる。

キーワード：エージェント、混合データベース、ルール抽出

1 はじめに

一般的に、数値を持つ属性と記号を持つ属性が混在しているデータベースは混合データベースと呼ばれている。様々な問題に対するシステムが整備されてくると、数値データベースや記号データベースだけでなく混合データベースの取扱いの必要性も同時に増加してくる。ここでは特に説明したい属性は数値属性であるが、その他の属性群は数値属性や記号属性が混在しているデータベースを対象とする。

データベースからの知識発見のプロセス(KDDプロセス)はFayyadら［1］でまとめられている。このプロセスの中にあるデータマイニングでは、与えられたデータベースの中からパターンを発見することが主なタスクである。特徴的なデータマイニング手法は文献［2]などでまとめられ、新たな手法の開発が続いている。その中に、探索的データ解析の代表にクラスター分析［3］［4］［2］がある。近年ではK-means法［5］から派生している様々な手法（［6］など)やエントロピー正則化法［7]、およびそこから派生した手法、そしてカーネル関数を用いた手法などが多く提案されている。従来から親しまれているウォード法［8］k-means法[5]など、クラスター分析に関する研究は、文献[3]に詳しくある。そこでは、与えられたデータ分布は、複数の分布モデルから生起した事象が混在した結果生じたものであるという仮定のもとでクラスター分析を行う手法もいくつか紹介されている。

いくつかの部分構造あるいはパターンが混在するような複雑なデータベースからそれらを抽出しようとするとき、クラスター分析は有効である。特徴属性を導く分類手法も多く提案されていて、それらの多くは始めから与えられているカテゴリーを重要視するが、得られるクラスタのラベルを生成するという意味［9］で本論文ではクラスター分析をベースにしたルール発見のアプローチをとることとする。そして、部分構造をさがしながら、その特性を説明する属性群もさがす問題を考える。

クラスター分析の多くの方法はデータを解析する前に、デザインパラメータや規範(評価関数)をあらかじめ与える必要がある。どういう与え方をすればいいのかという問題に対する回答は容易ではなく、それに対する研究は今なお続けられている[10]。また、分析を行うためには、クラスタリングに使用する属性もあらかじめ与える必要がある。クラスタリング属性は、個々属性の影響をできるだけ平等に考慮するため、なるべく互いに相関が少ないものを選択するのが望ましい[11]。あらかじめ与えられたクラスタリング属性がクラスタリング結果に大きく作用するので、得られた部分構造は実際の特性を表現できているのか、得られた結果をよく検討する必要がある。

データの部分構造を把握しようとするとき、クラスター分析を行いながら部分構造を記述する属性も選択することが望ましい。ところが、データ分布は、ときには次元をまたがって、複数の異なる部分構造から成るものであるとすると、クラスタリング属性を用いて、クラスター分析と同時にルールを記述する属性群を同定するのは注意を要する。例えば、ある部分構造は属

† An Agent-based Rule Extraction Method from Mixed Database
Mina RYOKE and Yoshiteru NAKAMORI

*1 筑波大学 大学院ビジネス科学研究科
Graduate School of Business Sciences, University of Tsukuba
*2 北陸先端科学技術大学院大学 知識科学研究科
School of Knowledge Science, Japan Advanced Institute of Systems and Technology

性ひとつだけで記述できるが、別の部分構造は、複数の属性を用いて記述するといった場合が考えられるからである。従来の多くのクラスター分析では、属性を変えながら分析を行うことが難しい。以上のことを考慮して、データベースからの部分構造を発見するエージェントの設計を提案することが本論文の目的である。

本論文の構成は次の通りである。2節で、本手法のアプローチを述べる。3節で、混合データベースを扱うための類似度について述べ、4節で、エージェントの設計、5節でエージェントの行動について述べる。6節で、本手法を用いた実験を述べたあと、7節で、まとめと今後の課題について述べる。

2 エージェント・アプローチ

本節では、ルール獲得における問題点を整理し、エージェントを用いる利点について述べる。

エージェント・アプローチは、現在ではさまざまな分野に適用されていて、それに応じて、定義自体もさまざまであると言える [12] [13] が、ここでは、多くとりあげられている、「内部に自身規範を持ちそれに従い行動するもの」をエージェントの定義とする。本論文では、ルールを獲得していくエージェントの内部規範設計を提案する。

知識発見手法に、遺伝的プログラミングなどの進化的アルゴリズムを用いた研究は多く見られる[14]。それらの多くは教師ありデータであり、予測精度を重要視している。コード表現を持つエージェントにより教師無しデータに対するクラスタリングを行いながら、属性を選択する手法も提案されているが、結論部分にモデルを持つルールを同定するものではない [15]。

ルール群の同定のために重要なことは、

- データ空間の部分構造をよくとらえること、
- 各ルールの特徴をよくとらえる属性を同定すること

等が、挙げられる。そこで、考えられる従来の方法のひとつは、部分構造をとらえるためにクラスタリングを行ない、部分集合ごとに条件部と結論部の同定を行なうことである。そのためには、次のステップが必要である。それは、

1. クラスタリング属性を選択し、
2. クラスタリングの評価規範を選択してクラスタリングを実行したのち、
3. 各部分集合について条件部属性の選択、結論部の形式の選択と結論部属性の選択

である。そして必要に応じて前ステップまたは前々ステップに戻り、繰り返す。

本論文では、より柔軟にデータベースの構造をとらえルールを抽出する目的のために、従来の多次元観測データに対して、前もってなんらかの数理モデルを仮定、あるいは評価関数を与えて解析をおこなうトップダウン式アプローチではなく、エージェントを用いたボトムアップ式アプローチをとる。

エージェントをデータ解析に適用することについて、他の利点を述べる。

- まず、後に定義する、各属性軸におけるデータ間の類似度と、全属性を用いたデータ間の類似度を用いることで、数値属性や記号属性、あるいはそれら両方を持つデータベースの解析が可能となる。
- そして、各属性軸で定義された類似度を用いて属性選択を解析プロセス中に行なうことができる。そして、属性選択によるオブジェクトのやりとりが可能となる。
- また、得られるルールの数は最終的に生き残るエージェントの数に相当するので、ルール数を明示的に与える必要がないことが挙げられる。
- さらに、エージェントが保持するオブジェクトにより、エージェントが持つルールの形態が変化する。

ファジィモデル [16] 同定にエージェントを用いたアプローチをとっている方法がある [17]。ここでは、全体として非線型な分布を持つデータに対し、エージェントは与えられた予測誤差以内の精度を持つ線形回帰モデルを同定していく。パラメータに直接精度を与えるので結果に大きく影響を与える。例えば、与えられたパラメータが予測誤差が小さいものとするとエージェントの数は多くなり、ひとつひとつのエージェントが持つ適用範囲が小さくなり、ブラックボックスとなってしまう危険を持っている。したがってパラメータの与え方が簡単ではない。また、属性選択は考慮されていない。

大局的な規範を与えずにクラスタリングをおこなうという試みは、文献 [18] でもなされている。しかしながら、これはトランザクションデータといった混合データベースに対して、ユーザーが与えるデザインパラメータにより2つの規範を使い分けるパラメトリッククラスタリングであり、我々のアプローチとはまったく異なる。

複数の規範を用いてクラスタリングを行う研究は多くなされている [19] [20]。特に、データ分布におけるデータの連続性と線型性を同時に考慮する方法は盛んに研究されている。文献 [20] では、部分構造を発見できたが、それぞれの特徴を記述する変数の選択方法はモデラーの知見を用いて、クラスタリングの後にあらためて検討している。本提案とは、クラスタリン

グ変数とモデルを作る変数の取り扱いが全クラスタのモデルに共通であるという点で異なっている。

3 類似度

エージェントの仕事について述べる前に、本節では、本論文で使用する類似度を定義する。

現在までに、さまざまな類似度が数多く提案されている[21] [3]。本論文では、類似度は、0から1の間の実数で類似度が高いほど大きい値をとるものとして次のシンプルな類似度を用いている。

データ空間の構造は、属性といくつかの属性で特徴づけられているオブジェクトからなるとする。ひとつの属性はデータ空間の次元に対応させる。オブジェクトの属性値は比例、間隔、順序、名義尺度のいずれで与えられていてもよいことにするため、それぞれの尺度ごとに属性値間類似度を定義する。

ここでは、各属性における類似度を属性値間類似度と呼び、全属性について求められた属性値間類似度を用いて、さらに、オブジェクト間類似度を定義する。ここで、集合 O をオブジェクト番号の添字集合、集合 X を属性番号の添字集合とし、

$$i,j\in O \qquad x\in X$$

そして z_{ix} をオブジェクト i の属性 x の値とする。

数値データを持っている属性 x における属性値間類似度 $Sim(z_{ix},z_{jx})$ は次のように定義する。

$$Sim(z_{ix},z_{jx})=1-\frac{|z_{ix}-z_{jx}|}{\max_{l\in O}\{z_{lx}\}-\min_{l\in O}\{z_{lx}\}} \tag{1}$$

記号データを持っている属性については、前述のように、様々な尺度を持つことが想定できる。従って、それに応じて属性 x における属性値間類似度 $Sim(z_{ix},z_{jx})$ は定義できる。

$$Sim(z_{ix},z_{jx})=\begin{cases}1 & \text{if} \quad z_{jx}=z_{ix}\\ 0 & \text{if} \quad z_{ix}\neq z_{jx}\end{cases} \tag{2}$$

もしくは、あらかじめ属性に対する知見がありその距離を別に同定することができるといった状況に応じて、各属性毎に類似度を定義することとする。

オブジェクト i と j の間の類似度 $SIM(i,j)$ は次のように定義される。

$$SIM(i,j)=\frac{1}{|X|}\sum_{x\in X}Sim(z_{ix},z_{jx}) \tag{3}$$

4 エージェントの設計

本節では、エージェントの内部について述べる。

まず、エージェントの代表オブジェクトについて述べる。どのエージェントも保有するオブジェクトの中から代表をもつ。それを、代表オブジェクトと呼ぶ。代表オブジェクトは、固定されているものではなく、エージェントが保有するオブジェクトの変化により置き換わる。エージェントが保有オブジェクトの数が2以下のときは代表オブジェクトの置き換えについてチェックしないが、エージェントが持つオブジェクトに変動があり、保有オブジェクトの数が3以上となると、代表オブジェクトの置き換えについてチェックを行い、エージェント内でオブジェクト間類似度の最大を持つペアに変化があれば、その都度、代表オブジェクトは入れ替わることとする。代表オブジェクトは、最もオブジェクト間類似度が高いペアのうちいずれか一方を、代表オブジェクトとしている。

次に、エージェントが同定する集合について述べる。それらを整理したものを次に述べる。()内部にあるのはその集合の名前である。

1．エージェントが持つオブジェクト集合(エージェントの視野) (View)
2．エージェントの特徴を示す属性集合(Char)
3．エージェントが回帰モデルをつくることができるか検討するために
   (a) 回帰属性の候補を示す属性集合(Survey)
   (b) 回帰属性の集合(Space)
4．エージェントがさらにオブジェクトを探すためのオブジェクト集合(Front)
5．さらに獲得される候補となったオブジェクト集合(FrontView)

以下に、それぞれの集合の詳細を述べる。

エージェント A は次式に従い、オブジェクトを取り込む。

$$View(A)=\{i\in O|SIM(i,i_A)\geq p_1\} \tag{4}$$

ここで、i_A は、エージェント A の代表オブジェクトである。また、パラメータ p_1 は、許容オブジェクト間類似度を示している。これは、そのエージェントが眺めることのできる視野範囲を示しており、集合 $View(A)$ は、エージェント A が持っているオブジェクト集合を意味する。

オブジェクト集合 $View(A)$ はどういう特徴を持つのかを探るため、属性集合 $Char(A)$ を次のように同定する。

$$Char(A)=\{x|\min_{\{i,j\in View(A)\}}Sim(z_{ix},z_{jx})\geq p_2\} \tag{5}$$

パラメータ p_1 はオブジェクト間類似度に関するもので

平均的なものであるが、ここで使われるパラメータ p_2 は属性値間類似度における同様の意味を持つものである。

属性集合 $Char(A)$ に属性 x が含まれているということは、属性 x 軸ではオブジェクト集合 $View(A)$ に含まれるオブジェクトが近くに存在することを示している。このことから、本集合は、エージェントがルールを作るときにその条件部を構成する属性を選択するための属性候補集合を意味する。

ある属性が属性集合 $Char(A)$ に含まれないとき、それは、その属性軸でオブジェクトがばらついていることを意味する。属性集合 $Char(A)$ に、目的属性が含まれ、他属性が含まれない場合や、どの属性も含まれない場合は、その場合は許容オブジェクト間類似度 p_1 やパラメータ p_2 が大きすぎることを意味する。そのときはエージェントがそれぞれのパラメータの設定を緩くする。

目的属性が数値属性で、ひろがりがあるとき、結論部は線形モデルを用いることができるかも知れないことを示唆している。そこで、次の集合を用いて調べる。

$$Survey(A)=\{k|\min_{\{i,j\in View(A)\}} Sim(z_{ik},z_{jk})\leq p_3\} \tag{6}$$

ある属性が属性集合 $Survey(A)$ に含まれるとき、それは、その属性ではばらつきがみられることを意味するので、次に、目的属性を線型回帰モデルで表現する可能性を検討する。そのため、他の属性と目的属性の相関を次のように調べる。まず、次のように目的属性値間類似度の最も小さなオブジェクトのペアを見つける。

$$(i_1,i_{|View(A)|})=\arg\min_{\{i,j)|i,j\in View(A),i<j\}} Sim(z_{iy},z_{jy}) \tag{7}$$

ここで $|View(A)|$ は、オブジェクト集合 $View(A)$ に含まれるオブジェクトの数を示す。

次に、オブジェクトの順番を与え、番号をつけなおす。

$$i_{k+1}=\arg\max_{\{i|i\notin i_1,\cdots,i_k\}} Sim(z_{iy},z_{iky}),\quad k=1,2,\cdots,|View(A)|-1 \tag{8}$$

このように、目的属性についてオブジェクト集合が並べ換えられる。その順番を用いて、数値属性を持つ各属性軸上では目的属性と相関が強いかを調べる。

$$Mon(x)=\frac{\sum_{l=2}^{|View(A)|-1} Sig(Sim(z_{il}x,,z_{i_1}x)-Sim(z_{i_{l+1}}x,z_{i_1}x))}{|View(A)-2|} \tag{9}$$

ここで、$Sig(x)$ は、次のように定義される。

$$Sig(x)=\begin{cases}+1 & \text{if } x>0\\ 0 & \text{if } x=0\\ -1 & \text{if } x<0\end{cases} \tag{10}$$

本論文では、より簡単なため、目的属性と各属性の間の単調性を調べて、次の説明属性集合を求めることとする。

$$Space(A)=\{x|Mon(x)\geq p_4,x\in X\} \tag{11}$$

属性集合 $Space(A)$ は、結論部における線型回帰モデルの説明変数候補集合である。現時点ではできるだけ容易に説明変数を探すために、属性集合 $Space(A)$ を定義し、含まれる属性を説明変数として使用しているが、説明変数間の相関をさらに検討することも考えられる。それについては今後の課題とする。

次に領域の拡張のための集合群を定義する。エージェント A は、次のオブジェクト集合 $Front(A)$ として同定する。

$$Front(A)=\{i\in View(A)\mid p_1<Sim(z_{ix},z_{i_Ax})\leq p_5,x\in Space(A)\} \tag{12}$$

エージェント A が線形回帰モデルを構成する属性空間で領域を拡大するために、現在の領域の端にあるオブジェクトの集合を同定している。これは、領域拡大の可能性をチェックするオブジェクトを決めていることを意味する。現段階では、少なくともひとつのオブジェクトが含まれるようにパラメータ p_5 は設定することとしている。

オブジェクト集合 $Front(A)$ に含まれるオブジェクトの視野に入るオブジェクト集合 $FrontView(A)$ は次により定義される。

$$FrontView(A)=\{i|Sim(z_{ix},\ z_{jx})\geq p_6|\ i\notin View(A),\ j\in Front(A)\} \tag{13}$$

パラメータ p_6 は、境界付近にいるオブジェクトからの視野を決定している。現段階では、パラメータ p_1 と同じ値をとるか、少なくともひとつのオブジェクトが含まれるような値としている。

オブジェクト集合 $FrontView(A)$ に含まれるオブジェクトについて、属性集合 $Space(A)$ を用いて作られた線形回帰モデルを用いて予測を行う。加えられたオブジェクトの2乗誤差平均が現在の2乗誤差平均よりも小さければ、そのオブジェクトは、オブジェクト集合 $View(A)$ に含まれ、パラメータ p_1 は、それを含むように小さくされる。ただし、属性集合 $Char(A)$ が空集合

になれば、パラメータ p_2 を小さくするか、たったいまパラメータ p_1 が小さくなる原因となったオブジェクトを追加することを止めて、パラメータ更新もしない。また、発見されたオブジェクトがまったく同じ属性集合 $Space(B)$ を持つエージェント B の $View(B)$ に含まれていたら、予測誤差が少ない方が奪い取る。$Space(A)$ と $Space(B)$ が等しくなければ、そのまま両方に属する。

5 エージェントの行動

5.1 ルール発見のプロセス

はじめに、許容オブジェクト間類似度 p_1（$0 < p_1 < 1$）が与えられる。エージェントが同定するオブジェクト集合や属性集合は、それぞれパラメータを必要とするが、本論文では、p_1 以外のパラメータはエージェントが変更している。例えば、オブジェクト集合 $FrontView$ に少なくとも1点はオブジェクトが含まれるように、パラメータ p_6 はエージェントにより変更される。そして、予測したい（数値）属性を指定する。

エージェントはデータベースに一つの代表オブジェクトと共に投入される。代表オブジェクトには、どのエージェントにも属していないオブジェクトが与えられる。そして先に定義した集合を同定して、保有するオブジェクトの数や同定した集合群に応じたルールを持つ。オブジェクトの数を増やせなくなったら、次のエージェントが代表オブジェクトと共に投入される。

投入された後は、エージェントは以下に述べるレベルに応じた行動をとり、領域を拡げるように大きくなっていく。データベースに投入されるエージェントの始めの状態はレベル2である。その後の過程によりそれぞれのエージェントの状態は異なったものとなる。どのオブジェクトもいずれかのエージェントにより保有されていて、エージェントが取り込む他のオブジェクトやエージェントをあらたに発見できないときや拡張できなくなったとき本プロセスは終了する。

領域の拡げ方には、2フェーズがある。フェーズ1では、エージェントはオブジェクトを獲得するように動き、そのときの指標は、代表オブジェクトとのオブジェクト間類似度と、先にのべたオブジェクト集合 $FrontView(A)$ による拡張である。フェーズ2では、エージェントは他のエージェントの領域を獲得するように動く。このときの指標は、それぞれのエージェントの代表オブジェクト間類似度と、同じくオブジェクト集合 $FrontView(A)$ による拡張である。

エージェントはオブジェクトあるいは他のエージェントを取り込み、必要であれば代表オブジェクトを置き換えていくが、代表オブジェクトの置き換わりにより、代表オブジェクトとのオブジェクト間類似度がパラメータ p_1 よりも小さくなるオブジェクトが出現する。そのようなオブジェクトは、エージェントから放出され、他のエージェントに取り込まれるか、そうでなければ、それ自体が代表オブジェクトとなってレベル2のエージェントとなり、他のオブジェクトをさがすことになる。

5.2 エージェントの行動レベルと得られるルール

次にエージェントの行動レベルについて述べる。各エージェントは、次に述べる各レベルに沿った動きをする。

レベル1

ここでは、データベースから、属性値間類似度 Sim とオブジェクト間類似度 SIM を導出する。記号属性ごとに何か知見があれば、記号を持つ属性の属性値間類似度の定義はその知見に従って決める。これは、全エージェントが同じ類似度テーブルを持つものとする。

レベル2

エージェント A は、オブジェクト集合 $View(A)$ の定義に従い、与えられたパラメータ p_1 より大きいオブジェクト間類似度を持つオブジェクトを取り込んでいき、属性集合 $Char(A)$ を同定する。このとき、次のルールが得られる。属性集合 $Char(A)$ に、目的属性とそれ以外の属性が含まれる場合のとき、

$$\text{if } \{x \text{ is } A_x | x \in \{Char(A) - y\}\}, \quad \text{then } y \text{ is } A_y \tag{14}$$

ここで y は、先に与えられている説明したい属性とする。A_x、A_y は、含まれるオブジェクトの数により、シングルトンか、区間か分布と形を変えて表現される。

属性集合 $Char(A)$ に、目的属性以外の属性が含まれる場合のとき、

$$\text{if } \{x \text{ is } A_x | x \in \{Char(A)\}\}, \quad \text{then } y \text{ is } A_y \tag{15}$$

このとき、目的属性の軸上でオブジェクト集合 $View(A)$ に含まれているオブジェクトは拡がっているので、後件部が線形モデルで表すことができる可能性がある。

エージェント A が、取り込むオブジェクトを発見できないとき、次のエージェント B がデータベースに、あらたな代表オブジェクトとともに放り込まれ、また取り込むオブジェクトを探す。

レベル3

どのオブジェクトもいずれかのエージェントに取り

込まれているとき、線形回帰モデル構築について、それぞれのエージェントは、属性集合 $Survey(A)$, 指標 $Mon(A)$, 属性集合 $Space(A)$ を同定して検討する。これは、レベル2からレベル3へ変化することができるかを検討することでもある。

属性集合 $Space(A)$ に、目的属性以外の数値属性が含まれているとき、エージェント A は回帰モデルを作ることができる。回帰モデルが同定されたとき、エージェント A は次のルールを持つ。

$$\text{if } \{x \text{ is } A_x | x \in \{Char(A)\}\},$$
$$\text{then } y = f(Expl(A)) \tag{16}$$

ただし、関数 $f(Expl(A))$ は、$Space(A)$ から目的属性をのぞいた属性の集合 $Expl(A)$ を説明変数とする回帰モデルとする。

エージェント A は、$Space(A)$ における回帰モデルに基づいて拡張をはかる。エージェント A は、オブジェクト集合 $Front(A)$ と $FrontView(A)$ を同定し、$Space(A)$ で張られる空間で拡張が可能かどうか調べる。$FrontView(A)$ に含まれるオブジェクトを、回帰モデルへ適用する。回帰モデル同定に使用されたオブジェクトによる平均2乗誤差と、新しく適用されたオブジェクトによる誤差を比較して、それが小さいなら、取り込み、拡張を行い、可能なかぎり繰り返す。

6 実験

ここでは、本手法の有効性を示すために、実験を行う。教師無しデータからのルール抽出問題では、得られた結果を評価することも容易ではない。そこで、ここでは非常に簡単でどういう結果が良いか直感的にわかる人工データを準備する。用いるデータを表1に示す。x_1 と x_2 は数値属性、x_3 は記号属性を持つ。ここでは、2つの実験を示す。両方の実験のいずれも、予測したい属性は x_2 とし、パラメータ p_1 は0.7とした。

本実験では、代表オブジェクトは類似度の高いペアのうちオブジェクト番号が若い方とする。

パラメータ p_1 以外は、各エージェント毎に値が異なる。許容オブジェクト間類似度 p_1 を与えると、視野は制限されるが、代表オブジェクトは実際にオブジェクトが存在する範囲を見るので、それは、p_1 より大きくなる。それを p_1^{AN} とする。AN は各エージェント毎に与えられる名前とする。現段階では、それをまず p_2 の初期値とし、各集合の定義とともに述べたパラメータに関する知見を用いて、探索的に各エージェントに決めさせている。よりよいパラメータの決定方法は今後の課題である。

実験1

図1に、表1の人工データに本手法を適用した結果を示す。数値属性と記号属性の両方を考慮した結果が得られた。得られたルールは次の通りである。

- 斜十字のクラスタ：条件部属性は、x_3, x_1 で、結論部は2つのオブジェクトからなる x_2 の区間
- 縦十字のクラスタ：条件部属性は、x_3 で、結論部は線形モデル $x_2 = 0.9090 + 3.085 \times x_1$
- 米印のクラスタ：条件部属性は、x_3, x_1 で、結論部は線形モデル $x_2 = 4.477 + 0.7593 \times x_1$

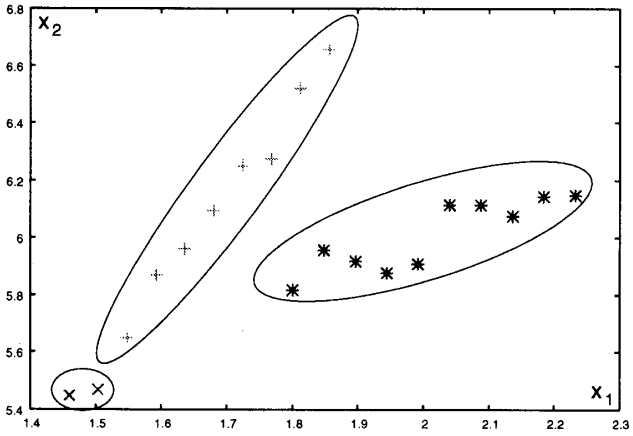


図1 Result in a mixed database.

表1 Artificial Dataset

Objects	Attributes x_1	x_2	x_3
1	1.460	5.448	II
2	1.504	5.471	II
3	1.548	5.649	I
4	1.592	5.871	I
5	1.636	5.962	I
6	1.680	6.090	I
7	1.724	6.250	I
8	1.768	6.275	I
9	1.812	6.520	I
10	1.856	6.658	I
11	1.800	5.817	II
12	1.848	5.957	II
13	1.896	5.918	II
14	1.944	5.876	II
15	1.992	5.909	II
16	2.040	6.114	II
17	2.088	6.114	II
18	2.136	6.076	II
19	2.184	6.144	II
20	2.232	6.150	II

実験2

実験1では、記号属性と数値属性が混在したデータを扱った。実験2では、エージェントの動きを示すために、オブジェクト間類似度が実験1よりも互いに似ているデータを扱う。このデータ分布では、0か1の属性値間類似度を持つ記号属性がひとつ含まれている3次元での実行よりも、数値属性のみの2次元での実行の方が、オブジェクト間類似度の値が似てくるため、終了までのエージェントの動作数が多い。そこで、数値属性のみ2次元のデータに本手法を適用しエージェントの様子を示す。

また、提案したエージェントの動きがわかりやすいように、あえてパラメータ p_1 を0.7という低い値にしている。この値が低いため許されるエージェントの視野が広くなり、後に示すように様々な状態を観察することができる。なお、この人工データではパラメータ p_1 をもう少し高い値にしても違う経過を経て同じ結果を得ることができる。

エージェントの様子は、表2から表10に示す。表中の1行目にある DN（Data Number の略記）の列はデータ番号を、RN（Representative Number の略記）の列はそのときの代表オブジェクト番号を、そして、AN（Agent Name の略記）の列はエージェントの番号を示している。

表2では、はじめにエージェント1は、代表オブジェクトを1として、オブジェクト2、オブジェクト3、そしてオブジェクト4を持っていることを示している。先に与えた許容オブジェクト間類似度から、これ以上領域を拡げられない。エージェント1は代表オブジェクトの移動についてチェックしたが、この状態ではオブジェクト1とオブジェクト2のオブジェクト間類似度が最も高いので移動はない。

次に代表オブジェクト5を持つエージェントが投入され、それは、はじめにオブジェクト4を見つけた。オブジェクト4は、既にエージェント1に保有されているが、エージェント1の代表オブジェクトとの類似度よりもエージェント2の代表オブジェクトとのオブジェクト間類似度が高いことから、オブジェクト4は、エージェント2に奪われる。そのあと、エージェント2は他にオブジェクト6を取り込み、表3のような、エージェント2が、代表オブジェクトを5とし、オブジェクト4とオブジェクト6を持った状態となった。

表3の後、エージェント2の保有するオブジェクト数が3を越えたため、エージェント2が代表オブジェクトを移動するかチェックしたところ、代表オブジェクトが5から4に変更した。表4では、エージェント2の代表オブジェクトが4となり、さらに保有するオブジェクトに6を追加した状態を示している。

表2 経過1

DN	RN	AN
1	1	1
2	1	1
3	1	1
4	1	1

表3 経過2

DN	RN	AN
1	1	1
2	1	1
3	1	1
4	5	2
5	5	2
6	5	2

様々な状態を見ることができるよう、許容オブジェクト間オブジェクト間類似度 p_1 を0.7と低めの値に設定しているため、表5では、エージェント2はさらに遠くのオブジェクトを取り込んでいる様子がわかる。

表4 経過3

DN	RN	AN
1	1	1
2	1	1
3	4	2
4	4	2
5	4	2
6	4	2

表5 経過4

DN	RN	AN
1	1	1
2	1	1
3	4	2
4	4	2
5	4	2
6	4	2
11	4	2
12	4	2
13	4	2

そして、エージェント2は、代表をオブジェクト12へと変更した（表6）。その結果、許容オブジェクトの制限から、オブジェクト3を放出し、さらに拡張と代表オブジェクトの変更を繰り返し、表7では、さらにテリトリーを拡げた様子を示している。表7では、オブジェクト3とオブジェクト4、オブジェクト9とオブジェクト10以外はエージェント1かエージェント2に属していることになっている。

その後、新しくエージェント3が投入されたが、他からオブジェクトをとることができずに、一つのオブジェクトしか持たないエージェントができた様子を示している。また新しくエージェント4が投入され、オブジェクト5を取りこんだ。そしてさらにエージェント5が投入され、オブジェクト9とオブジェクト10を取り込んでいる。表8はそのときの状態を示している。ここで、いったんどのオブジェクトもいずれかのエージェントに保有されていることになる。

ここから、表9の状態になるには、次のような経緯がある。まず、エージェント5がオブジェクト8を発見し、エージェント2の代表オブジェクトとの類似度よりもエージェント5の代表オブジェクトとの類似度

表6 経過5

DN	RN	AN
1	1	1
2	1	1
4	12	2
5	12	2
6	12	2
11	12	2
2	12	12
13	12	2

表7 経過6

DN	RN	AN
1	1	1
2	1	1
5	16	2
6	16	2
7	16	2
8	16	2
11	16	2
12	16	2
13	16	2
14	16	2
15	16	2
16	16	2
17	16	2
18	16	2
19	16	2
20	16	2

が高いので、オブジェクト8は、エージェント5が保有することとなる。次にエージェント5は、オブジェクト7を発見し、同様に類似度を比較して取り込む。エージェントは、保有するオブジェクトが3以上になれば、代表オブジェクトを置き換えるか常にチェックすることとしているが、エージェント5がオブジェクト7を取り込んだとき、代表オブジェクトがオブジェクト7となった。さらに、エージェントは、オブジェクト6を発見し、今度は新しいエージェント5の代表オブジェクト7との類似度とエージェント2の代表オブジェクトとの類似度を比較して、オブジェクト6を取り込んだ。

エージェント2とエージェント5はそれぞれ*Front*を同定し、領域拡張を図るがここでは拡張できなかった。特にエージェント5は拡張できそうだが、代表オブジェクト6から離れた端のオブジェクトを*Front*(5)にはオブジェクト10が含まれたため、そこから得られる*FrontView*(5)には、オブジェクト11しか発見できなかった。

次にエージェントがオブジェクトを探すのではなく、他のエージェントを取り込むようになる。このとき代表オブジェクト間類似度を用いて相手を探す。このときの許容オブジェクト間類似度は、本実験では先と同じく0.7を用いているが、エージェント同士が広い範囲を探索することができるように、より小さくしてもよい。その結果、エージェント3がエージェント4を取り込んだ。エージェント3は、オブジェクト3、オブジェクト4とオブジェクト5を保有することとなった。*Front*(3)を同定し、*FrontView*(3)には、オブジェクト1とオブジェクト2が含まれた。エージェント3が持

表8 経過7

DN	RN	AN
1	1	1
2	1	1
3	3	3
4	4	4
5	4	4
6	16	2
7	16	2
8	16	2
9	9	5
10	9	5
11	16	2
12	16	2
13	16	2
14	16	2
15	16	2
16	16	2
17	16	2
18	16	2
19	16	2
20	16	2

表9 経過8

DN	RN	AN
1	3	3
2	3	3
3	3	3
4	3	3
5	3	3
6	7	5
7	7	5
8	7	5
9	7	5
10	7	5
11	16	2
12	16	2
13	16	2
14	16	2
15	16	2
16	16	2
17	16	2
18	16	2
19	16	2
20	16	2

つデータを用いて同定された回帰モデルにオブジェクト1とオブジェクト2を適用して誤差をもとめたところ、エージェント3が持つデータを適用して得られた平均2乗誤差よりもどちらも小さいので両方ともエージェント3に取り込まれた。そして、表9の状態が得られた。

続いて、エージェントは*Front*を同定し領域拡張を図る。この状態で*Front*を作ることができたのは、まず、エージェント2であった。*Front*(2)に属したのは、オブジェクト11であった。0.7の視野を持っているため、*FrontView*(2)には、3、4、5、6、7、8と大量にオブジェクトがあったが、そのときエージェント2が持つ回帰モデルに当てはめたところ平均2乗誤差が大きかったので、とりいれられなかった。他のエージェントも領域拡張にはいたらなかった。

代表オブジェクト間類似度からエージェント3はエージェント5を見つけ、全部奪った。これ以上どのエージェントも領域拡張や他の取り込むべきエージェントを発見することができず、最終結果は表10となった。

図2に、最終結果を示す。

また、最後まで残ったエージェントは、条件属性はx_1で、それぞれが持つ回帰モデルは次の通りである。

- エージェント3（十字）：$x_2=4.477+0.7593\times x_1$
- エージェント2（×印）：$x_2=0.8517+3.117\times x_1$

表10 経過9

DN	RN	AN
1	4	3
2	4	3
3	4	3
4	4	3
5	4	3
6	4	3
7	4	3
8	4	3
9	4	3
10	4	3
11	16	2
12	16	2
13	16	2
14	16	2
15	16	2
16	16	2
17	16	2
18	16	2
19	16	2
20	16	2

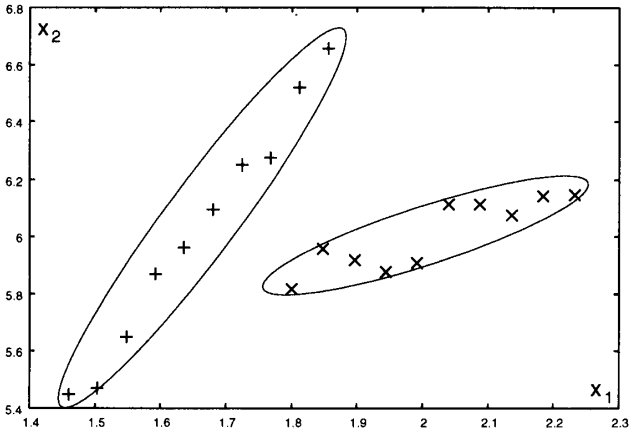


図2 Result in a numerical database

7 まとめと今後の課題

本論文では、混合データベースからの知識発見のためのエージェントの設計について提案した。混合データベースに対応するために属性値間類似度とオブジェクト間類似度を用いた。本論文では0から1の間の実数値をとる類似度を用いたが、エージェントは他の類似度を用いることも可能である。

提案されたエージェントは、はじめは類似度のみを用いてオブジェクトの獲得、放出をおこない、次第にオブジェクトの数が増え、また特性となる属性を複数の集合群から発見すると回帰モデルを構築し、それに従いオブジェクトの獲得や放出をおこなう。

今後の課題は、より多くの次元を持つデータベースに適用し、各エージェントによるパラメータのよりよい設定法と属性集合 *Space*(*A*) に含まれる説明属性候補からの互いの相関を考慮した簡便な説明属性選択法などがあげられ、さらに改良をおこなう。また、得られたモデル群に基づいたルールベース型モデルの構築へと発展させていく。

参考文献

[1] U. M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, Vol. 17, No. 3, pp. 37-54 (1996)

[2] J. Han and M.Kamer, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers (2001)

[3] B. S. Everitt, "Cluster Analysis", Third Edition, Edward Arnold (1993)

[4] R. O. Duda, P. E. Hart and D. G. Stork, "Pattern Classification", Wiley-Interscience Publication (2001)

[5] J. MacQueen,"Some Methods for Classification and Analysis of multivariate Observations",Proc. of the 5th Berkeley Symp. Math. Statist, Prob., Vol. 1, pp. 281-296 (1967)

[6] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms",Prenum (1981)

[7] S. Miyamoto and M. Mukaidono, "Fuzzy c-means as a regularization and maximum entropy approach", Proc. of the 7th International Fuzzy Systems Association World Congress (IFSA'97), June 25-30, Vol. 2, pp. 86-92 (1997)

[8] J. H. Ward, Jr., "Hierarchical grouping to optimize an objective function", "J. Amer. Statist. Assoc", Vol. 58, pp. 236-244 (1963)

[9] D. H. Fisher, JR., M. J. Pazzani, and P. Langley (eds), "Concept Formation: Knowledge and Experience in Unsurpervised Learning", Morgan Kaufmann Publishers, Inc. San Mateo, California (1991)

[10] C. Fraley and A. E. Raftery: "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis", The Computer Journal, Vol. 41, No. 8, pp. 578-588, (1998)

[11] 中森義輝・領家美奈, "ファジィモデリング再考", 日本ファジィ学会誌, Vol. 5, No. 3, pp.453-464 (1993)

[12] J. M. Epstein and R. Axtell, "Growing Artificial Societies : Social Science from the Bottom Up", The MIT Press, Cambridge, Massachusetts (1996) (服部正太・木村香代子(訳) : 人工社会 複雑系とマルチエージェント・シミュレーション, 共立出版社,1999)

[13] R. Axelrod, M. D. Cohen, "Harnessing Complexity",Perseus Books Group (2001) (高木晴夫(監訳), 寺野隆雄(訳), 複雑系組織論, ダイヤモンド社,2003)

[14] A. A. Freitas, A survey of Evalutionary Algorith-

m for Data Mining and Knowledge Discovery, in: Ghosh, A.: Tsutsui, S. (Eds.) Advances in evolutionary computation. Springer-Verlag., (2001)
[15] Y. Kim, W. N. Street and F. Menzcer: Feature Selection in Unsupervised Learning via Evolutionary Search, Proc. of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2000)
[16] T. Takagi and M. Sugeno,"Fuzzy Identification of systems and its applications to modeling and control", IEEE Trans. on Syst., Man and Cybern., SMC-15, No. 1,pp. 116-132 (1985)
[17] T. Ma and Y. Nakamori, Agent-based approach to Fuzzy Modeling, IFIP Workshop Group 7.6 Workshop on Virtual Environments for Advanced Modeling, pp. 34-35, (2002)
[18] G. D. Ramkumar and A. Swami, "Clustering Data without Distance Functions", Data Enginnering Bulletin, Vol. 21, No. 1, pp. 9-14 (1998)
[19] R. N. Dave, "An Adaptive Fuzzy c-Elliptotype Clustering Algorithm", Proc. of NAFIPS 90: Quater Century of Fuzziness, Vol. I,pp. 9-12, (1990)
[20] 領家美奈・中森義輝, "適応型ファジィ回帰のパラメータ設定法", 日本ファジィ学会誌, Vol. 10,No. 2, pp. 330-337 (1998)
[21] B. R. Boyce, C. T. Meadow and D. H. Kraft, "Measurement in Information Science", Library and Information Science Series, Academic Press (1994)

(2003年7月31日 受付)
(2003年11月6日 再受付)
(2003年12月12日 採録)

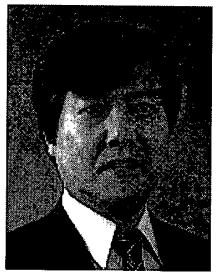
[連絡先]
〒112-0012 東京都文京区大塚3-29-1
筑波大学大学院ビジネス科学研究科
領家 美奈
TEL：03-3942-6889
FAX：03-3942-6829
e-mail：ryoke@gssm.otsuka.tsukuba.ac.jp

著者紹介

領家 美奈 [正会員]

筑波大学大学院ビジネス科学研究科

1998年大阪大学大学院基礎工学研究科物理系専攻制御工学分野博士後期課程修了。博士(工学)。1998年4月より2002年9月まで北陸先端科学技術大学院大学助手。2002年10月から筑波大学大学院ビジネス科学研究科講師。この間2001年6月より2002年5月まで国際応用システム解析研究所研究員。1993年度日本ファジィ学会論文賞、1999年度システム制御情報学会学会賞奨励賞受賞。日本ファジィ学会、システム制御情報学会、環境科学会会員。



中森 義輝 [正会員]

北陸先端科学技術大学院大学知識科学研究科

1979年京都大学大学院工学研究科数理工学専攻博士課程修了。工学博士。甲南大学理学部応用数学科勤務を経て、1998年4月より北陸先端科学技術大学院大学教授。1984年10月より1985年11月まで、国際応用システム解析研究所研究員。1986年4月より環境庁国立環境研究所客員研究員。1992年9月より大連理工大学客員教授兼務。日本ファジィ学会、計測自動制御学会、環境科学会、IEEEなどの会員。

An Agent-based Rule Extraction Method from Mixed Database

by

Mina RYOKE and Yoshiteru NAKAMORI

Abstract：

This paper proposes an agent-based rule extraction method from mixed database. In the database, some heterogeneous structures are often included together. In addition, some symbolic attributes and some numerical attributes are often stored together in the databases. The proposed agent has a role for collecting data objects and detecting a local rule, based on a similarity and identified sets. The interaction between the agents focuses on squabbling about objects. % considering attributes. The agent identifies an if-then rule based on the collected data. The consequence of the obtained rule is changed by the agent, depending on the condition.

Keywords： Agent, mixed database, rule extraction

Contact Address：**Mina RYOKE**
Graduate School of Business Sciences, University of Tsukuba
3-29-1 Otsuka, Bunkyou-ku, Tokyo 112-0012
TEL ：03-3942-6889
FAX ：03-3942-6829
E-mail ：ryoke@gssm.otsuka.tsukuba.ac.jp