論　文

事象ごとの生起確率から未知事象発見を支援する手法とそのアンケート調査への適用

Discovery-Aiding of Unknown Events from Event-Probabilities, Applied to Questionaires

砂山　渡*　大澤　幸生*　谷内田　正彦*
Wataru Sunayama　Yukio Ohsawa　Masahiko Yachida

* 大阪大学大学院基礎工学研究科システム人間系専攻
Dept. Systems and Human Science, Osaka University, Toyonaka City, Osaka 560-8531, Japan.

1998年6月26日　受理

Keywords: discovery of unknown events, questionaire analysis.

Summary

The real world includes various unknown elements, objects, phenomena, and so on. Causes of phenomena to be observed is difficult to specify, because they might be uncoverable for the observer. That is, it takes too much time to consider all the possible causes, or sometimes they are even unable to be considered because they are unknown. Now then, we aim at constructing a learning method by which to learn unknown causal-links and unknown basic causes, i.e., deeper causes than those assumable as hypotheses. A major contribution of this paper is that the input data are not combinations of simultaneous occurence of events, but the average occurence of each event given as probabilities. By determining the position of new nodes in the probabilitic network, the user reasons for him/herself, what the new node means in the real world. The system is applied to analyzing data obtained by a few questionaires.

1. はじめに

人間が努力してもなかなか得られない科学上の原理を発掘する研究が，データ・マイニングや人工知能の分野において近年盛んに行なわれている [Washio 97]. 実際，それまで知識に入っていなかった重要な未知の事象を発見することは，どの時代においても社会現象や自然現象を理解するための中心的な課題である.

しかし，自然や社会は非常に複雑に構成されていて，目の前に現れるいろいろな現象の原因となる未知の事象が一体何であるかを記述するのに必要な知識もまた書き尽くし難く，機械に与えられている限られた知識ベースで表現するのはきわめて困難となる.

そこで本論文では，未知事象の発見を自動化するのではなく，未知事象が何であるか人間が理解するのを支援するシステムを考える．特に，アンケート調査から未知の大衆心理を発見するという問題を取り上げ，本システムの原理を示し，効果を実験により検証する.

以下，**2**章には従来のデータ解析手法を示し，本研究の目指すことを明らかにする．**3**章では我々の扱うデータ解析の対象の範囲と，その表現方法を述べる．**4**章では構築した未知事象発見支援システムの概要を，**5**章でその詳細を示す．**6**章では実験によるシステム評価を行ない，**7**章で結論を述べる.

2. 従来のデータ解析手法と本研究の動機

新しい因子を発見することを考えると，論理に基づく帰納学習において新述語を生成する枠組み [Muggleton 88] がある．しかし，そこでは主に既知の知識の記述単位（アトム）を組み合わせた新しい記述の単位が生成されるのであり，知識中で全く考慮されていなかった未知要因を推定するという目的はむしろ因子分析 [柳

井 90] の目的に近い.

因子分析では，潜在因子と呼ばれる，直接の観測が不能な原因を入力データから推定する*1．しかし，因子（変量）の値の組合せ（以下，**組合せデータ**と呼ぶ）で与えられるデータを多数集めて入力とするところに実用上の問題があった．アンケート調査でいえば，各回答者が各質問にどう答えたかという組合せデータ（回答者１人に１つの組合せデータを得たもの）を

回答者 A：経済改革が重要，福祉改革は不要
回答者 B：経済改革が不要，福祉改革も不要
回答者 C：経済改革が重要，福祉改革も重要
…

と与えたり，各組合せデータの得られる確率として

経済改革が重要，福祉改革も重要：30%
経済改革が重要，福祉改革は不要：10%
経済改革が不要，福祉改革は重要：40%
経済改革が不要，福祉改革も不要：20%

などと与えるのである．Bayesian Network において事象間の新しい因果関係を見いだしたり [Heckerman 95]，新しいノードを生成したりする [Geiger 96] 学習もあるが，やはり組合せデータの集合が入力である．

しかし，実際にどのようなデータが入力できるかは非常に重要な問題であり，次のようなことが考えられる．

1. アンケートの調査者と解析者が異なる．
2. アンケートで組合せデータが得られない．
3. N 個の変数について，2^N 個のデータが得られることは滅多にない．

この１について，例えば，ある政党が政策決定を行なう際に新聞に載せられた世論調査の結果を分析する場合も考えられる．アンケート調査者とアンケート解析者が同一とは限らない場合，得られるデータは上記のような組合せデータではなくて全員の回答を平均した確率データであることが多く，

経済改革が重要: 40%, 福祉改革が重要: 70%

のように各質問項目に対して１つずつだけとなる．

このような平均データであれば，膨大な組合せデータよりも格段に少ない量のデータであるため，様々な情報源から容易に入手できる．2については，人によって答えた質問項目が異なっている（上の例で福祉のことは分からないので無回答とする）場合が考えられ，回答者の感情（**6・1 節**参照）を考えた場合，組合せデータよりも各質問項目ごとの平均データを集めるほうが信頼できる場合もある．

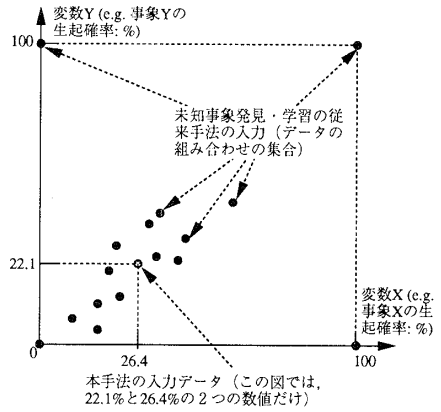


図１　本研究で扱う，「**各事象の平均の**」確率データ

3については，例えば，10 個の二値変数について 100 の組合せデータが与えられた場合，$2^{10} = 1024$ 通りの変数の取り得る値の組合せパターンについてそれぞれ正確な確率を得るにはデータ数が不十分である．一方，各変数については 100 個のデータがあり，この各変数ごとの確率のほうが，各組合せパターンの確率値よりも精度は高いといえる．

これらの理由により，図1のような組合せデータよりも乏しいデータをもとに未知事象を発見する枠組みが必要となる．Bayesian Network によって推論 [Heckerman 94] したり，Dempster-Shafer の理論によって事象の生起確率を求めること [上野 87] はこのようなデータでも可能だが，このような乏しいデータから未知事象の学習を行なう手法は従来なかった．

Bayesian Network 上で，事象ごとの確率からモンテカルロ法によって組合せデータを大量に生成する方法もある [Henrion 88] が，乏しいデータに加えて新しい情報を考慮するわけではないので，新しいノードを追加するという，本来豊富なデータが必要な学習の目的には用いられない．しかし，上述のように，組合せデータではなく，事象ごとの平均の生起確率だけを入力として未知事象を発見することは現実の問題として重要であるから，本研究ではあえてこの問題に挑む．

しかしながら，事象ごとの平均の生起確率という乏しいデータから未知原因を知るためには，学習時の強いバイアス [山田 97] がなければ学習結果が決まらない．そこで我々は，初めに知識ベースとして既知の事象間の関係を備えておき，それがある程度信頼できるという仮定（後述の**仮定 1**）をこのバイアスとして用いる．

*1　特に探索的因子分析では，潜在因子の数も未知であるから，未知事象の存在そのものを計算で求めようとする意味をもっている．

初めに知識を備えておく考えは，因子分析においても潜在変数間の関係や，潜在変数と観測可能な変数の関係を全て構造として与えておく共分散構造解析で採用されている [柳井 90]．しかし，共分散構造解析は，未知事象同士・未知事象と既知事象の関係についての仮説が予め立てられている場合にしか用いることができない．これに対して，我々の問題では知識中の事象と事象間の関係だけを既知とし，この使える構造だけを用いて未知事象間・未知事象と既知事象の間の関係を推定するのである．

3. 事象と論理モデル

3・1 事象の論理関係

我々は事象間の関係を全て，条件がアトムの連言であるルール（**AND** ルールと呼び，式 (2) で表す．“, ” は AND を表す）または選言となっているルール（**OR** ルールと呼び，式 (1) で表す）で記述し，このように記述できる事象[*2]間の関係（これを以下，事象間の**論理関係**と呼ぶ）のみを考える．

$$Y \leftarrow Y_1 \text{ or } Y_2 \text{ or } ... \text{ or } Y_m. \quad (1)$$

$$Y_1 \leftarrow X_1, X_2, ..., X_n, C. \quad (2)$$

$$inc \leftarrow Z_1, Z_2, ..., Z_n. \quad (3)$$

既知の事象間の論理関係に，次の仮定を設ける．

> **仮定 1：** 知識ベースは，世界の事象の論理関係をできる限り網羅するつもりで構築される．

以下，式 (1), (2), (3) の意味を述べる．今，ある事象 Y を考える．事象 Y を，Y の原因別に $Y_1, Y_2, ..., Y_m$ という別名で呼ぶことにすると，式 (1) のような OR ルールとなる．すなわち，式 (1) は「Y は，$Y_1, Y_2, ..., Y_m$ のどれかが真である場合，かつその場合に限って真である」を意味する．ここでは，仮定 1 によって Y の原因を m 個に特定していることになる．このような $Y_1, Y_2, ..., Y_m$ のうち，特定の 1 つ Y_1 が式 (2) で定義される．すなわち Y_1 は，$X_1, X_2, ..., X_n$ が全て真である場合のみ，かつその場合確率 $P(C)$ で真となる事象である．ここでも，仮定 1 によって Y_1 の生起する条件が n 個に特定される．式 (3) は，$Z_1, Z_2, ..., Z_n$ が矛盾する，すなわち同時には真にならないという制約である．

これらのルールを，事象の論理関係のグラフで図 2 のように表現する．このグラフは，有向無循環のグラフ（Directed Acyclic Graph, 以下 **DAG** と略す [Pearl 93]）で，ノードが事象，リンク（枝）が論理関係を表す．また，ノード X はそのノードが表す事象 X が起こる確率値 $P(X)$ を属性値としてもつ．これが **1 章**で述べた「事象ごとの」平均の確率データである．

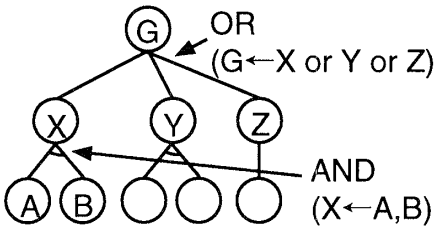


図 2 DAG の例．全てのノード（アトムに相当）は新ノードと OR または AND のいずれかの関係でリンクされる

3・2 この論理モデルの含む範囲

次のような事象や事象の集まりが，上のモデルの対象となる．

1) **真理値が 0 か 1 に固定された事象：** 確率値が 0（偽）か 1（真）の事象として扱う．
2) **離散多値の事象：** 離散二値（0 か 1）ではなく，離散多値の真理値を取り得る事象を扱う場合には，取り得る値のそれぞれに関してノードを 1 つずつ用意し，それらのノード間の矛盾（式 (3)）を予め与えることによって記述できる．

一方，連続値（例えば水の流量など）は取り扱わないが，必要ならば定性化によって，この枠組に収められる．

なお，因果関係の全てを記述するのが式 (1) から (3) のような論理関係だという主張は行なわない．実際，因果関係の表現方法がまだ確立されていない [Glymour 95] ので，本論文でも因果関係ではなく**論理関係**という言葉を用いている．後述するように本論文は，自然現象としての因果律に限らず，事象間の関係を仮定 1 を前提として表した論理関係と，観測データとの**ずれ**（**4 章**）を，やはり論理関係によって表現しようとする試みである．そして，ここで論理という表現方法（論理表現 [坂原 85]）を用いる理由は，人間にこのずれを理解させるという我々の目的にとって，論理表現が計算機から人間に提示する情報として単純で理解しやすい（「最も」理解しやすいと主張する根拠はないが）と考えられるからである．

4. 未知事象発見支援システムの構成

1 章に述べたように，本研究で構築した**未知事象発**

*2 「事象の真理値」とはいわないのが一般的であるが，ここでは簡単のため「事象」とその生起を表すアトム（真ならば事象が生起しているとする）を同じ記号で表す．

見支援システムの目的は，未知原因の正体を人間であるユーザが知るためのヒントを与えることである．そもそも，人間が新事実を発見するのは，普段とは異なる現象に着目したときが多い．そして，初めにもっていた知識と，実際に起こっている観測事象とのずれに着目して知識を更新するのである．この発見を支援するのが，未知事象発見支援システムの役割である．

本システムのタスクは，次の入力から出力を得ることである．

入力：各事象ごとの個別の生起確率と知識構造　従来から，確率値を入力とする因子分析は多く研究されているが，**1章**に述べたように，本論文では多くの因子（変量）の値の組合せ，すなわちデータの組合せデータが得られない場合を扱うのがポイントである．また，**3・1節の仮定1**を前提とした論理関係が既知事象間に与えられる．

出力：新ノードと新リンク　未知の事象を表す，もとの知識ベースに含まれなかったノードと，知識ベース中の事象との論理関係を表す．この未知原因は計算機の知識ベースには含まれていないものであるから，計算機にはその正体（実世界において，新仮説が何を表すか）を知ることはできない．

その正体について知っている可能性があるのは，計算機よりむしろ独自に豊かな経験をもっている人間であろう．未知事象発見支援システムの出力をユーザが見て，未知原因の正体を考え，理解することがこのシステムの目的の全てである．しかし，この結果の解釈を個人によって行なう場合には主観が強く解釈に影響する可能性がある．できるだけ多くの人の意見をもとにして客観的な解釈を得ることが望まれる．

入力からいかにして出力を得るかという手続きについて簡単に述べると，図3のようになる．まず，確率値のわかっていない事象（前述の**3・1節**で，確率値の与えられないアトム）の確率値を補う（**5・1節**で後述）．しかし，知識ベース中の事象間に未知の，すなわち知識ベース外の要因のせいで依存が生じている可能性もある．そこで，知識ベースだけから期待される各事象の生起確率と，実際の確率値とのずれをもとに，既知事象間の依存関係を計算する（**5・2節**で後述）．そして，それらの依存のもとである**未知原因**と既知の事象との論理関係を求め（**5・3節**で後述），これをユーザに対して表示する．

以上のプロセスについて，図3の流れに沿って，**5章**で順に説明していく．

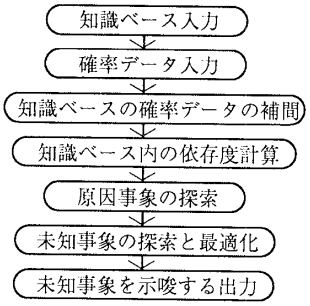


図3　未知事象発見支援システムの流れ

5. 未知事象発見支援システムの詳細

5・1 確率データの補間

知識ネットワーク上の事象で，確率データ（**3・1節**のノードの属性値）が与えられていないものがあったとする．このような場合，データの得られている事象の確率値をもとに，確率データの抜けている事象の確率値を補間する．

具体的には，知識ネットワーク上の確率値に関して，**エントロピー最大化**の高速手法 [Cheeseman 83] を用いる．エントロピー最大化とは，知識ベース中で独立とみなされている事象が，できるだけ独立に近くなるように各事象の生起確率を求める操作である．これは，もとの知識ベースの完成の度合いをできるだけ信頼して，**仮定1**に基づいて欠けている知識が最小となるように補間するという意味をもっている．例えば，図2においてノード Xの確率値が欠けていたとすると，Xを結論とするANDルールの条件 A, Bと，Xを条件として G を結論とするORルールの条件の事象 Y, Zの確率値を元に，A, B, Y, Zが独立に最も近くなるように Xの確率値を補間する．

5・2 事象間の依存

式 (2)，(1) の各ルールにおける条件事象の間には，実際には正の依存，独立，負の依存のいずれかの依存関係がある．

まず，**正の依存**とは，事象間の共通原因などのせいで複数の事象が同時発生しやすくなる関係をさす．正の依存が最も強いのは，事象間に包含関係が成立する場合である．すなわち，ANDルールでは最も確率の小さい条件事象がその他全ての条件事象に含まれる場合であり，ORルールでは最も確率の大きい事象がその他全ての条件事象を含む場合である．このときのルールの結論部の事象の確率値がそれぞれ，ANDルールにおける結論の上限値（式 (4)）とORルールにおける結論の下限値（式 (9)）とになる．

逆に**負の依存**とは，互いに同時生起しにくい排他的な事象間の関係のことである．負の依存が最も強いのは，事象生起が互いに矛盾（排反）する場合であり，AND, ORルールとも，条件事象のいずれかが起こる確率が1を超えない範囲で，条件事象の全てが起こる確率が最小となる．このときのルールの結論部の事象の確率値がそれぞれ，ORルールにおける結論の上限値（排反事象の選言の確率はその和で与えられることから式(7)）とANDルールにおける結論の下限値（ANDルールとORルールは互いに双対関係にあることから式(7)で各事象の確率の補数をとって，式(6)）になる[Bonissone 87]．

以上のように，各ルールの結論部の確率の上限と下限が条件部の確率から与えられるのであるが，条件部の事象に一切依存がなければ，すなわち独立であれば，結論の確率は上限と下限の中間の値（AND, ORの独立の定義式から式(5), (8)）をとる．

ANDルール： $A \leftarrow X_1, X_2, ..., X_n$

$$\overline{P(A)} = \min\{P(X_1), P(X_2), \cdots, P(X_n)\} \tag{4}$$

$$P(A) = \prod_{i=1}^{n} P(X_i) \tag{5}$$

$$\underline{P(A)} = \max\left\{0, \sum_{i=1}^{n} P(X_i) - n + 1\right\} \tag{6}$$

ORルール： $A \leftarrow X_1$ or X_2 or ... or X_n

$$\overline{P(A)} = \min\left\{1, \sum_{i=1}^{n} P(X_i)\right\} \tag{7}$$

$$P(A) = 1 - \prod_{i=1}^{n}(1 - P(X_i)) \tag{8}$$

$$\underline{P(A)} = \max\{P(X_1), P(X_2), \cdots, P(X_n)\} \tag{9}$$

以上のように与えられる確率値の上下限の値と，実際の結論部の事象の確率値から，ルールの条件事象間の依存関係とその依存の強さ（次の依存度）を計算することができる．その計算手続きを以下に示す．

〔1〕 **事象間の依存度の計算**

依存度とは，知識ベース中の各アトムが表す事象の生起の間の依存関係である．ルール r の依存度 $depend(r)$ は，ルールの条件の事象が互いに独立時の結論の確率と実際の確率との差を，排反で-1，包含で1の値をとるように，その結論の確率の上限と下限の大きさから正規化したものである．

すなわち，式（4）から（9）の値を用いてルール r の条件事象が互いに独立なときの結論 A の確率値を a，A の上限と下限の確率値をそれぞれ$\overline{P(A)}$と$\underline{P(A)}$，実際の確率値を $P(A)$ と表したとき，依存度を式（10）から（13）のように与える．

ANDルール： $A \leftarrow X_1, X_2, ..., X_n$

$$\text{正の依存}：depend(r) = \frac{P(A) - a}{\overline{P(A)} - a} \tag{10}$$

$$\text{負の依存}：depend(r) = -\frac{P(A) - a}{\underline{P(A)} - a} \tag{11}$$

ORルール： $A \leftarrow X_1$ or X_2 or ... or X_n

$$\text{正の依存}：depend(r) = -\frac{P(A) - a}{\underline{P(A)} - a} \tag{12}$$

$$\text{負の依存}：depend(r) = \frac{P(A) - a}{\overline{P(A)} - a} \tag{13}$$

したがって，未知（実存するのに知識ベースに含まれない）事象がない場合には，知識ベースで条件事象が互いに独立とされている全てのルールの依存度は0となる．そこで，条件事象が互いに独立とされている，言い換えれば，条件事象の間に共通原因や矛盾関係がない全てのルール[*3]について，式（10）から（13）の $|P(A) - a|$ の値がデータの誤差（〔2〕項の値）を超えるルール r に関して，原因の探索を行なう．

〔2〕 **誤差への対策**

本システムにおける依存度の値の誤差は，ルールの後件部確率のとり得る範囲に影響する．わずかな誤差でも，ルールの結論の確率が独立時の値からずれると，ありもしない依存を検出してしまうからである．

我々はこの点を考慮して，ある閾値以下の依存は無視する．すなわち，ルール r の結論の事象 A の確率 $P(A)$ の想定される誤差（95 %の有意水準での信頼区間の幅[白旗 92]）e に対して，式(10), (13)では $|depend(r)|$ が $e/(\overline{P(A)} - a)$ を，式(11), (12)では $|depend(r)|$ が $e/(a - \underline{P(A)})$ を超えるときにのみ，それぞれ該当する（正または負の）依存を検出する．これによって，誤差によって誤った依存をとらえないようにした．ここで確率値 p に付随する誤差 e は，各ノードの事象が二項分布に従うとすると，

$$e = 1.96\sqrt{\frac{p(1-p)}{\text{標本（回答者）数}}} \tag{14}$$

と与えられる．例えば，**6・1節**の実験例では30人からのデータを，**6・2節**の実験例では，15万人からのデータを用いたので，観測誤差はそれぞれ最大で $e = 18\%, 0.3\%$

*3 事象同士に直接の因果関係がある場合も，そのうち1つ以上の事象を共通原因とみなすことでこの中に含める．

と計算され，実際には20%と5%とを依存検出の閾値とした．

5・3 未知事象の探索

最後に本システムは，アトム間の未知の依存を新仮説を表す新ノードという形にまとめて出力する．

〔1〕 **強論理関係による絞り込み**

1つの新仮説が知識ベース中のどの事象の共通原因として表れるかは，知識ベースの規模の指数オーダの可能性がある（新仮説が知識中のどのアトムの原因であるかという組合せになるから）．したがって，知識の重要な部分に焦点を当てた処理が必要となる．我々は，この焦点を**強論理関係**によって当てる．

強論理関係とは，あるルール r（例えば $Y \leftarrow X_1, X_2$ とする）において，rの結論部 (Y) と，Yの生起を大きく左右する条件（X_1 または X_2のいずれか）との間のリンクのことをさす．すなわち，ANDルールにおいては，結論と，確率の低い条件とのリンクであり，ORルールにおいては，逆に確率値の大きい条件とのリンクとなる．例えば，部屋の電灯がつくためには電力が供給されていて，ブレーカーが落ちておらず，スイッチが入っていなければならない（3条件のANDルールによる結合）．すると，電灯がついているかどうかを強く左右する原因は，『電力が供給されている』や，『ブレーカーが落ちていない』など普段は満たされている事象ではなく，『スイッチが入っている』という，他の事象に比べて確率の小さい事象となる．また，夜に部屋が明るいためには，『部屋の電灯がついている』か『懐中電灯がついている』が必要だとすると（2条件のORルールによる結合），前者の条件のほうが確率が高いので『夜，部屋が明るい』にとって重要な条件となる．

我々は，未知原因の知識ネットワーク中での位置を，ここで述べた論理関係の強さに応じて限定する．具体的にはANDルールにおいては結論の確率の半分以下の確率値をもつ条件事象と結論事象とのリンク，ORルールにおいては結論の確率の半分以上の確率値をもつ条件事象と結論事象とのリンクとして定めた．この事象間の強い関係を表すリンク（**強論理リンク**と呼ぶ）を出力時に明示し，未知原因は，強論理リンク上のノードに隣接する位置にのみ生成する．すなわち各依存ルールの結論事象を起点として，ネットワークの結論→条件の向きに強論理リンクを辿って，その上のノード間にのみ共通原因となる未知事象を生成するようにする．

〔2〕 **事象間の依存を説明する未知ノードの生成**

予想される確率値からのずれから，事象間の依存度を求める方法についてここまでに述べた．この依存を説明する未知事象としては，図4に示した次の3種類を考える．

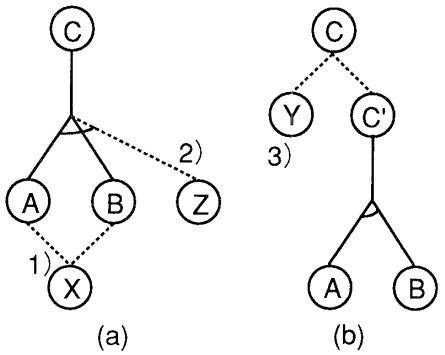


図4 未知事象の種類

3種類の未知事象

1) 事象 A, Bの背後にさらに根本的な原因となる新仮説 Xが存在（Xと A あるいは Bが矛盾するような関係も含む．Aと Xが等価であれば，単純な Aと B の間の依存となる）．

2) 事象 A, Bの他の未知原因 Z が A, Bと並列に存在．

3) 事象 Cに A, B以外の原因が存在．

新ノードが「A, Bより深い原因のレベル」である1)，「A, Bと同レベル」である2)，「A, Bと Cの中間レベル」である3)のほかに，「Cと同レベル以上となるケースが考えられるが，このケースは Cを条件とするルールでの2)または3)のケースとして含まれる．

全てのノードの確率値から，各事象間の依存を1), 2), 3)のいずれであるか判別する．すなわち，

- **依存度の絶対値が1を超えた場合** ANDルールの結論が下限値を下回れば2)，ORルールの結論が上限を上回れば3)と判定する．
- **依存度の絶対値が1以下の場合**
  - **依存度の絶対値が0の場合** Aと Bが独立であるとする．なお，これ以外の場合，$C \leftarrow A, B$ を**依存ルール**と呼ぶ．
  - **依存度の絶対値が負の場合** 1)のうち Aと Bが矛盾するケースか，それとも2)のケースかは，A, B, Cの値だけからは区別できない（Cが下限より小さければ2)となるので区別できる）．そこで，結果の解釈の時点で1)と3)のケースの区別を必要に応じて人間が行なうものとする．
  - **依存度の絶対値が正の場合** 1)のケースか3)のケースかは，A, B, Cの値だけからは区別できない（Cが上限を超えれば2)となるので区別

できる）．そこで，上の場合と同様に結果の解釈の時点で人が区別する．

〔3〕 未知事象の探索

知識ネットワークに，上述のように求めた依存のケースに応じて新たな未知事象のノード，または事象間の矛盾のリンクを付加するのであるが，ここで付加するノードの数を最小化する．これは，付加するノードがもとの知識ベースにおいて欠落していた程度を最小化することに当たり，先の**仮定1**の通り，もとの知識をできるだけ信頼する，すなわち不足している知識（ここでは生成されるべき新ノード）が最小となるように未知事象を推定するためである．未知事象探索の手続きは次の通りである．

未知事象探索アルゴリズム

STEP1. 依存ルールの結論を導く条件事象を強論理リンクの範囲内（先述の〔1〕項）で全て集める（集合 S とする）．

STEP2. 全ての依存ルールの条件事象をそれぞれ2つ以上ずつ導く S の部分集合のうちで，要素数が最小の部分集合を探索する（この探索には遺伝的アルゴリズムを用いる）．

STEP3. 選ばれた条件事象間に未知事象を表す新ノードを付加する．

例えば，2つのルール $X \leftarrow P,Q$ と $Y \leftarrow U,V$ が依存ルールであった場合，P,Q,U,V を結論にもつルールの条件事象を集める（STEP1）．それらが強論理関係によって，それぞれ $\{A,B\},\{D,E\},\{B,C\},\{E,F\}$ が取りだされたとすると $S=\{A,B,C,D,E,F\}$ となり，STEP2の制約を満たす最適な部分集合は $\{B,E\}$ となる（STEP2）．この2つの事象 B と E の共通原因として未知事象を付加する（STEP3）．

〔4〕 制約違反の解消

事象の確率値が式 (4) から (9) の上限・下限の制約を違反する場合，この制約違反も依存と同じく，既存の知識ベースが完全でなく，実際に存在する条件が全てノードとして与えられていないことに起因する（先述の〔2〕項参照）．

そこで，AND（OR）ルールの結論の確率が下限（上限）を違反する場合には，そのルールもしくは，そのルールの条件を導くルールに新たな条件を補うことによって下限値を下げる（上限値を上げる）．この操作は，AND（OR）ルールが強論理リンク上のリンクを含む場合にのみ行なわれる．そして，この制約違反の解消のための条件も新ノードとして生成される．

6. 実　　験

本システムの有効性を示すために，以下の実験を行なった．動作環境はSUN SPARC Station10GX(64MB)であり，プログラムはC言語で書かれている．

6・1 アンケートからの大衆心理の発見 1

好きな動物がどんな特徴をもつ動物かを調べる，以下のアンケートを行なった．

まず初めに，5種類の動物に関する知識ベース（図5のノード P,Q とそれらの隣接枝以外のグラフ*4）を，被験者となる30人の意見をもとに構築した．それらは，動物の特徴を表す命題「猫はかわいい」,「犬は賢い」と，動物の特徴から好きな動物を導くルール

猫が好き ← かわいい動物が好き，猫はかわいい．

（かわいい動物を好きな人が，猫をかわいいと思っていればその人は猫が好きだ）である．

次に，2つの質問に答えてもらうアンケートを行なった．アンケートの方法は5種類の動物と，7種類の特徴からの選択方式で，複数回答が認められている．

アンケート1：好きな動物は何ですか．

回答：犬 (83.3%)，リス (63.3%)，猫 (60.0%)，あらいぐま (56.7%)，ペンギン (53.3%)，...

アンケート2：どんな特徴の動物が好きですか．

回答：かわいい（83.3%），賢い（66.7%），小さい（63.3%），足が速い（53.3%），...

被験者は30人であるが，「被験者Aは犬が好きで，賢い動物が好きで，大きい動物が好きで...」などという，各個人が各質問にどのように答えたかという組合

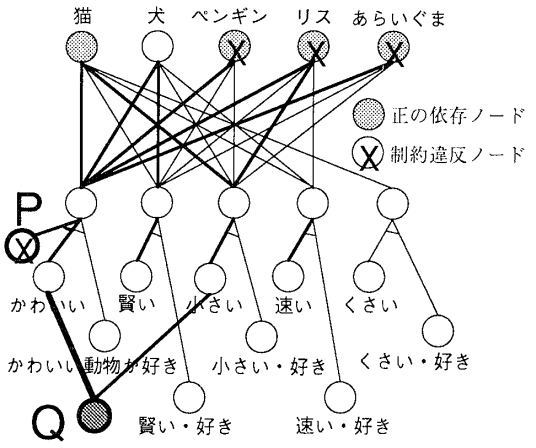


図5 好きな動物の特徴

*4 図5の「かわいい」のノードは見やすさのために縮約したノードで，実際は「猫はかわいい」,「犬はかわいい」など用意された命題の数のノードがある．

せデータは残されない．すなわち，システムの入力データは「犬が好きな人：83.3%，賢い動物が好きな人：66.7%...」などの，各事象ごとの確率値である．これは，被験者が各自のデータが残ることを嫌うという一般的な性質（名前を残さない場合でも，被験者は，自分の組合せデータから個人名が推定されること自体を嫌うことが多い）による要請である．このような被験者の感情（回答へのはじらいなどまで含む）を考慮することは，被験者が正直に回答するためにはきわめて重要である．このように組合せデータが入力されないという点で，**1章**に述べたように入力データが乏しいことになる．

知識ベース自体が被験者達の意見によって構築されたのであるから，上記の5種類の動物と，7種類の特徴を関係づける知識については，先の仮定1が成り立っていると考えられる．したがって，期待されるのは，これら以外の特徴が入力データをもとに新ノードとして生成されることである．この問題では本システムの処理は，以下のようになった．

確率データの補間　図5で名前のついていない中間ノード（ANDとORを別々に扱うため自動的に付加されたノード）の確率を補間する（**5・1節**）．

依存の計算　（**5・2節**）与えられた全ルールの条件事象間の依存を計算した結果，図5の「猫」,「ペンギン」,「リス」,「あらいぐま」ノードを結論とする各ルールの条件事象間に，それぞれ強い正の依存が見いだされた．

未知事象の探索　（**5・3節**）計算された4つの正の依存をたった1つで同時に説明する図5の新ノードQが,「小さい」と「かわいい」の間の正の依存を表すノードとして付加された（未知事象の個数の最小化）．最後に,「ペンギン」,「リス」,「あらいぐま（ノード）」を結論とするルールが，それぞれ結論の確率が確率値制約の下限を下回る違反を犯したので，この制約違反を解消するノードPが付加された．

図5の新ノードQは,「好きな動物の原因として,『かわいい』に近い感情が重要である．その感情によって，小さい動物が好まれやすい.」と解釈できる．実際，人間は赤ん坊に近い形や行動をとる動物を好むことが多く，上の解釈はこのような癖をとらえた結果といえる．

しかし，注目すべきもう1つの重要なポイントは，新ノードPが,「かわいい」の条件として新たに付加されたことである．そのままこの出力をよめば,「かわいい動物が好きな人でも，かわいくてかつ，何らかの条件（ノードPに当たる）を満たさなければ好きには至らない」ということになるが,「何らかの条件」とは何であろうか．システムはそれ以上詳しい知識をもっていないから，この問いかけをするまでの役割しか果たさない．ここから先は，人間が自分の経験や他の人との議論を頼りに考えるのである．

この場合，我々は3名の議論によって,「かわいい」は非常に主観的な特徴であり，動物と，その動物を好む人の個性によって，どのようにかわいいかが分かれるのだという結論を得た．すなわち,「かわいい」といっても,（動物，人）のペアによってその特有のかわいさがあり，この「特有の」が「何らかの条件」だという解釈である.「かわいい」が主観的で，明示的な定義のない概念であることを人に気づかせ，このような議論から結論を得たのは，まさに未知事象理解の支援という本システムの狙い通りであった．

6・2　アンケートからの大衆心理の発見2

大規模なアンケートではデータが膨大になるため，各回答者の組合せデータではなく，各質問項目に対する回答の確率データだけが残るか公開される．我々はその公開された確率データのみをもとに，アンケート調査者との関わりをもつことなく大衆心理の解析を行なった．ここにその一例を示す．

1996年の衆議院総選挙に関連して行なわれた世論調査の結果から，国民が時事問題に対して，互いにどのような相関関係をもっているかを調べる実験を行なった．世論調査の手法は，調査対象者を電話番号簿から無作為抽出した成人男女15万人として，質問紙に基づく電話聴取方式である[日経リサーチ 96]．

まず，これまでと同じ形式の知識ベース（式（15），（16）がそれぞれ，国民の支持と，政党の制約に関する知識となる）を与える．

$$lib \leftarrow lpr. \tag{15}$$

（「自民党を支持する（lib）のは自民党の公約のいずれかに興味がある（lpr）場合である.」）

$$lpr \leftarrow lwf \textbf{ or } lecn \textbf{ or } lgyo \textbf{ or } ltax \textbf{ or } ledu. \tag{16}$$

（「自民党の政策には，福祉lwf，経済$lecn$，行政$lgyo$，税制$ltax$，教育$ledu$関連のものがある.」）

次に，与える入力データは，世論調査から得られた各党を支持する人の割合（$P(lib) = 0.307$など）および，それぞれの種類の公約に関心がある人の割合（$P(lwf) = 0.025, P(lecn) = 0.039, P(lgyo) = 0.015$, …）の確率データである．これらはやはり，組合せデー

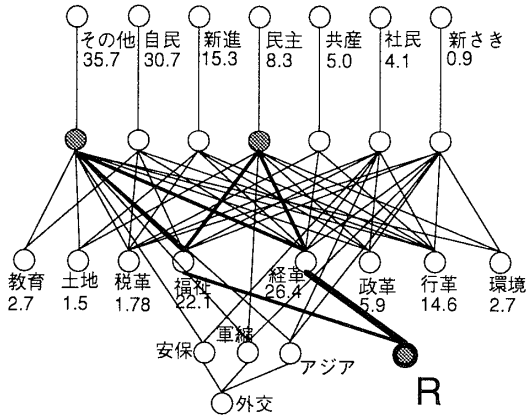


図6　支持政党と政策の関係

タではない．プロセスは先のアンケートと同様であるから詳細を述べないが，この場合は知識ベース中のノードの約2/3の確率データが欠けていたので，先に述べたデータの補間を行なった．このデータを，本システムにかけた結果を 図6に示す．

まず，太い線で示された強論理関係が絞り込まれた．その上で，民主党支持者と，その他の支持政党不明（ノード）に強い正の依存が認められ，それらを説明する新ノードとして，経済改革と福祉の間に正の依存を示唆するノード Rが付加された．ノード Rの意味として，「政党支持の決定原因として，経済改革と関連の深い生活水準の維持，向上に当たるような要因が重要であり，それが福祉など他の要因にも関連，影響を及ぼしている.」ことが示唆される．

ここから後は，先のアンケートと同様に，人間が経験によって深い理解を行なう．今回は，11人が意見を出して議論した結果，未曽有の経済危機を抱える中を高齢化に突入する我が国において，国民の日常生活が圧迫され，そこから行政への不信や福祉への期待が生じているのだという解釈にまとまった．この結果は，有権者の危機感を汲み取ったものといえよう．

7. 結　　論

各人の組合せデータの得られないアンケートの結果から，データに隠れた未知の原因および論理関係に関して，人間が理解するための示唆を与えるシステムを構築した．ここでは，組合せデータがないことによる情報量の乏しさを，もとの知識ベースが対象とする世界をほぼ網羅しているという仮定1によって補っている．

今後はマーケティングにおいて，消費者へのアンケートから消費者の中に大衆心理として潜んでいる新製品への未知のニーズを把握するなどの応用を筆者らは考えている．

謝　　辞

本研究を進めるに当り，貴重な議論をして頂いた鷲尾隆助教授（大阪大学産業科学研究所），滝寛和氏（三菱電機（株）），寺野隆雄教授（筑波大学）に感謝します．また，査読者のコメントは表現の改善にとって非常に有効であったので，ここに記して感謝します．

◇ 参 考 文 献 ◇

[Bonissone 87] Bonissone, P.P., et al.: A Layered Architecture for Reasoning with Uncertainty, *Proc. International Joint Conference of Artificial Intelligence (IJCAI'87)* pp. 891 – 898 (1987).

[Charniak 91] Charniak, E.: Bayesian Network without tears, *AI Magazine*, Winter, pp. 50 – 63 (1991).

[Cheeseman 83] Cheeseman, P.: A Method of Computing Generalized Bayesian Probability Values for Expert Systems, *Proc. IJCAI'83* pp.198 – 202 (1983).

[Geiger 96] Geiger, D., et al.: Asymptotic Model Selection for Directed Networks with Hidden Variables, *Technical Report of Microsoft Research (MSR-TR)* 96-07 (1996).

[Glymour 95] Glymour, C.: Available Technology for Discovering Causal Models, Building Bayes Nets, and Selecting Predictors:The TETRAD II Program, *Proc. Knowledge Discovery and Datamining (KDD'95)*, pp.130 – 135 (1995).

[Heckerman 94] Heckerman, D.: Causal Independence for Probability Assessment and Inference Using Bayesian Networks, *MSR-TR* 94-08 (1994).

[Heckerman 95] Heckerman, D.: A Tutorial on Learning with Bayesian Networks, *MSR-TR* 95-06, Microsoft Research (1995).

[Henrion 88] Henrion, M.: Propagating Uncertainty in Bayesian Networks by Logic Sampling, *Uncertainty in Artificial Intelligence 2* (Lemmer, J.F and Kanal, L.N. eds.), North-Holland (1988).

[Muggleton 88] Muggleton, S.: A strategy for constructing new predicates in first order logic, Proc. of the Third European Working Session on Learning, Glasgow. Pitman. (1988).

[日経リサーチ 96] 日経リサーチ：世論調査, http://www.nikkei.co.jp/hensei/election/seron.html, 10/10 – 10/13 (1996).

[Pearl 93] Pearl, J.: Aspects of Graphical Models Connected With Causality, Proceedings of the 49th Session of the International Statistical Institute, Italy, Tome LV, Book 1, Florence, pp. 399 – 401 (1993).

[坂原 85] 坂原茂: 日常言語の推論，東京大学出版会 (1985).

[白旗 92] 白旗慎吾：統計解析入門（共立出版）pp.127 – 129 (1992).

[上野 87] 上野晴樹，石塚 満 共編：知識の表現と利用，オーム社 (1987).

[Washio 97] Washio, T. and Motoda,H.: Discovering Admissible Models of Complex Systems Based on Scale-Types and Identity Constraints, *Proc. IJCAI'97* pp.810–817 (1997).

[山田 97] 山田誠二：適応エージェント，共立出版 (1997).

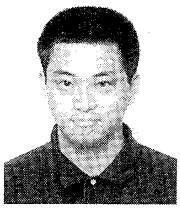
[柳井 90] 柳井晴夫他：因子分析—その理論と方法—, 朝倉書店 (1990).

〔担当委員：小林重信〕

著 者 紹 介

砂山　渡(学生会員)

1995年大阪大学基礎工学部制御工学科卒業．1997年同大大学院博士前期課程修了．現在，同大大学院博士後期課程在学中．知識の発見や精錬に関する研究に興味を持つ．
〈sunayama@yachi-lab.sys.es.osaka-u.ac.jp〉



大澤　幸生(正会員)

1990年東京大学工学部電子卒業．1995年同大学院博士課程修了．博士（工学）．現在，大阪大学基礎工学部助手．1994,1998年度人工知能学会全国大会優秀論文賞．IEEE，AAAI，情報処理学会各会員．　〈osawa@sys.es.osaka-u.ac.jp〉

谷内田　正彦(正会員)

1971年大阪大学大学院工学研究科修士課程修了．同年同大基礎工学部制御工学科助手．同助教授を経て同学部情報工学科教授，1994年同学部システム工学科教授，現在に至る．工学博士．画像処理，人工知能，移動ロボットなどの研究に従事．著書「ロボットビジョン」(昭晃堂),「コンピュータビジョン」(丸善，編著）など．情報処理学会，ロボット学会など各会員．〈yachida@sys.es.osaka-u.ac.jp〉