仮説推論機構を用いた多変量データ 解析支援システム

南 弘 征*

北海道大学工学部

水 田 正 弘

情報図形科学講座

佐 藤 義 治

A Multivariate Data Analysis Supporting System with Assumption-based Reasoning Function

Hiroyuki MINAMI, Masahiro MIZUTA and Yoshiharu SATO

要 旨 データ解析システムは、ユーザインタフェースの整備など、より使いやすくなってきている反面、初心者が誤用する可能性も大きくなってきている。このような問題に対し、著者らはこれまでに知識処理による支援機能を有したデータ解析システムを提案したが、用いている推論方法のために処理実行過程や効率面について問題を残していた。

本論文では、データ解析作業に知識処理を応用するための考察を行ない、解析作業に対し問題解決の枠組を適用し、さらに推論機構の一部に非単調推論の一実現形式である TMS を導入することで、より効率的な推論方法を考慮したシステムの提案を行なう。また、具体的な実行例を通じて、従来のシステムとの動作の違いを比較検討し、本システムがより有効であることを示す。

1. はじめに

近年、S 言語 (Becker *et al.*, 1988) などのデータ解析ソフトウェアにおいて、処理内容や機能面においてさまざまな拡張、発展が見られる。これらは、ワークステーションにおける X-Window, パーソナルコンピュータにおける Windows などのグラフィカルな操作環境にも対応することで極めて高い操作性を提供している。このようなパッケージの充実はデータ解析作業の効率を上げ、誤操作を減じる効果の他、利用者層をより広げる効果にもつながっていると考えられる。

しかし同時に、操作性のよいパッケージは統計学やデータ解析に関して不慣れなユーザが利用した場合、時として手法の誤用などを引き起こすことになりかねない。このような弊害に対するアプローチの一つとして、データ解析に関するサポートシステムの研究が 1980 年代から行なわれ

*現在小樽商科大学 商学部 社会情報学科

Keywords: Knowledge engineering, Statistical software

ている (Gale, 1986 など)。これらの研究例で報告されているシステムは、特定の手法における実行過程に対して細かな支援を行うものと複数手法から解析すべきデータおよび解析者からの情報などを通じて適切な手法を示唆するものに大別できる。しかし、データ解析に関する知識をほとんど持たないユーザの場合に対しては、その両方の機能を有するシステムが求められる。

著者らは、解析作業過程を知識工学という問題解決プロセスとして定式化し、解析手順列を示唆し実行するようなシステムの構築例を報告している (Minami *et al.*, 1993)。しかし従来の例では探索方法を Prolog 処理系に依存したため、形式的には正しいが内容を考えると不自然あるいは冗長な解析作業を示唆、実行することがあった。特に、処理列の再試行を行なう際には推論機構の制約により逐次後戻りを行なっていたため、専門家からみれば再試行に至った原因が明確な場合であっても、最前の処理から順にやり直す必要があるなど、主として知識処理の面で課題を残していた。

以上のような背景に基づき、本論文ではデータ解析作業についてできるだけ多くの部分をサポートすることを念頭に、知識処理に関して考察を行ない、解析の実行過程に関する支援において非単調推論の一実現形式である TMS (Truth Maintenance System, のちに Reasoning Maintenance System と改名) (Doyle, 1979) を用いたシステムの構築例について報告する。

2. データ解析における知識処理

データ解析と知識処理を融合した研究は、二つに大別される。一方は、定性的に知識処理の枠組に統計学の概念を導入して柔軟な処理を実現するもので、ベイズ確率を用いたものや Shafer-Dempster 理論を用いたものなどがある (古川・溝口 (編), 1988 など)。他方、本論文などが取り扱っているものは知識処理を応用してデータ解析作業に役立てようというものである。

後者の報告例として代表的なものに、CAI (Computer Assisted Instruction) 利用を念頭に置いて開発された回帰分析支援システム REX (Gale, 1986) がある。REX は、回帰分析で用いる計算式、指標値やその判断基準など、統計家が作成した戦略を、システム内部で保持している。ユーザが観測値のデータセット名を入力すると、戦略に従い種々の指標値を計算し、その値に応じ必要な場合はユーザに判断を求めた上で、解析を続行する。

また他に、重回帰分析の各過程を細分して階層化し、それらの過程毎に生じる中間データや指標値、行われる計算処理を、それぞれオブジェクト指向におけるクラスやインスタンス、メソッドとして取り込み、オブジェクトの階層を下りることで解析作業を進めていく支援システムとして、RASS (中野ほか, 1991) がある。これらは、(重) 回帰分析という単一手法を対象としている。

多くの解析手法の中から、ユーザの解析目的に合わせた的確な手法を選択するような報告例の一つに、Statistical Navigator (Brent, 1988) がある。これは、データの特性、外的基準の有無、解析目的などを入力していくことで、適切な手法を示唆するシステムである。内部では、個々の属性に与えられた重みの計算によって、最終的に適用手法を一つに絞り込んでいくという戦略が用いられている。そのため、実際の解析作業でしばしば用いられる複数手法の組み合わせまでは考慮されていない。また、回帰分析用として Smalltalk 環境で構築された DINDE (Oldford & Peters, 1988) などが他の手法への適用や複数手法の組み合わせについて実現可能性を示しているが、具体的な報告例は少ない。

限定された手法、用途での報告例に留まっている理由の一つとして、知識処理で用いることのできる知識の特徴と、データ解析で用いられている知識との対応づけが容易ではないことを挙げることができる。これについて、以下で説明する。

3. 解析作業と知識処理との対応

一般にエキスパートシステムなどの知識処理では、特定の問題領域を対象とし、主としてその領域に関する定性的な知識ベースが用いられる。しかし、データ解析作業は本来、定量的な情報を取り扱うものであり知識処理の導入にはさまざまな考察が必要となる。

3.1. 知識内容

データ解析において用いられる知識は、個体-項目型、類似度/非類似度といった正方行列などのデータ構造や個々の項目の尺度や単位など Metadata と呼ばれる「形式的知識」と、原データの実験計画、測定方法、解析結果に対応する解釈のための情報などの「領域知識」の二つに大別できる。ただし、完全に二分されるわけではなく、特に測定単位などは、両者の性格を合わせ持つ情報として、取り扱いが重要視されている (Hand, 1993)。

定性的という観点から考えるならば、領域知識を取り扱う知識支援は比較的容易と思われる。しかしながら、データ解析作業はその性格上、分野を問わずきわめて基本的な作業であるため、問題領域を特定することができない。一方、現在の知識工学では、全ての分野に関する領域知識を同一の枠組で実装することは事実上不可能である。従って、汎用性を考慮するならば、解析支援システムでは形式的知識を取り扱わざるを得ない。

しかし仮にある処理に関するあらゆる指標値を知識として盛り込むことができて、得られた値の解釈には領域知識を用いなければならない。領域知識が欠落した解析作業は単なる数値処理に過ぎず、意味ある解析とするためにはシステムを利用するユーザが領域知識を補うことが前提となる。

このような考察も含めてデータ解析に関するエキスパートシステムの作成は不可能であるとする批判的意見もあるが、REX, Statistical Navigator などの研究例が相応の評価を受けていることから、他分野におけるエキスパートシステム同様、解析作業に対する知識処理の応用は少なからず有効であると思われる。しかし、利用できる知識が限定されている以上、他分野でのエキスパートシステムで想定されているような作業の自動化などは困難であり、現在では、ある程度用途を限定し、明らかな誤りを起こさないような (less stupid) システムの構築が現実的とされている (Streitberg, 1988)。

3.2. 問題の定式化

知識処理の適用のためには、解析作業そのものを定式化しなければならない。著者らは、解析内容を決定づけるために、解析対象となるデータの形式的情報と解析目的をユーザからの入力とした上で、図 1 に示すような定式化を試みている。これに関する詳細は Minami *et al.* (1993) を参照されたい。

ところで、システムで有している知識だけを基に、観測値と解析目的から処理内容を一意に限定してしまうことは望ましいこととはいえない。むしろ、はじめから複数の処理列についての適

図1. 「処理」過程の定式化

用可能性を前提とし、結果によっては他の処理列の実行も考慮するような柔軟な枠組が求められる。

また、解析処理列を構築する際には、データの具体的な値に言及することなく、適用可能な処理を形式的知識を用いることにより、ある程度選択できる部分と、例えば回帰分析における多重共線性など、具体的な（指標）値に言及することで判断を確実に行なうことができる場合も考えられる。支援する解析手法を限定した場合には後者だけを考慮すればよいが、手法の選択に際しては前者によるところも少なくない。

そこで本システムでは、手法選択・処理列構築の段階までは定性的な情報のみを用い、処理結果に関する再検討の段階で、定量的判断を含む具体的処理を行なうこととした。しかし、既に述べた通り、知識処理の枠内において定量的判断を柔軟に行なうためには、従来用いてきた推論方法には直接適用しにくいとため、以下で考察を加える。

3.3. 推論方法

図1に示した定式化のうち、観測データから中心処理、中心処理結果から出力方法に至る処理列を構築する過程は人工知能の分野における問題解決プロセスとみなすことができる。

ここでいう問題解決とは、初期状態と目標状態が与えられた際、初期状態から目標状態を達成するような作用子の列を構築するプロセスである。作用子は自分が適用される状態（前提条件）と、自分が適用されることで失われる状態（削除リスト）と加わる状態（付加リスト）で構成されている。データ解析作業においては、データに関する情報が状態述語となるほか、解析処理が作用子となり、それぞれの処理について前提条件や削除・付加リストを用意することになる。また、初期状態には観測データが、目標状態としては解析結果が該当することになる。解析結果は未知であり、あらかじめ与えることはできないため、Minami *et al.* (1993) では、結果出力を行なう処理を状態述語を消失させる作用子と定義した。これによって、空である状態を目標状態と置くことが可能となる。

問題解決を行なう際に基本的な探索方針となる推論方法は、与えられた初期状態を変化させて

いくことで終了状態への到達を試みる（いわゆる cut & try に近い）前向き推論と、終了状態から逆に初期状態へ到達することを試みる後向き推論の二つに大別される。後向き推論は目標状態から探索を開始するため目標状態に至る可能性のある初期状態の数が少なければ、特に指針を持たずに探索を行なう前向き推論に比して無駄な探索を行なわないことが期待される。

しかし、後向き推論をデータ解析作業での問題解決に適用することは難しい。理由として、先に触れた通り定量的な情報が主である解析作業において判断基準として何らかの指標値が必要な場合、実際に処理を実行する以前、つまり具体的な出力値が得られていない状態では、推論で用いる指標値も得られないこと、本来未知である目標状態から逆に観測データへ至る処理列を構築するのが事実上不可能であることなどが挙げられる。

従って、データ解析作業への適用を行なう場合は前向き推論を選択することが妥当と考えられ、著者らもこれまで Prolog 処理系を用いた前向き推論を利用してきた。しかし、単純な前向き推論では失敗に至った履歴を利用できないため、情報のフィードバックが得にくい。

概念を図で示すと、Minami *et al.* (1993) で用いられているのは図2の(A)に相当する。しかし、実際の解析作業、特に探索的データ解析では解析過程において出力結果に即した試行錯誤が行なわれることを考えると、フィードバックを得にくい枠組は実際の作業手順と比べ不自然な過程を強いることになる。可能であれば、失敗に至った原因に即したところまで戻り必要な修正を行なったのち再試行できる、(B)のような枠組が望ましい。

図2. 処理列試行の概念図
(円が処理を，実線の矢印がデータの流れを，
点線の矢印は再試行の順序を表す)

そこで本論文で報告するシステムにおいては、形式的知識に基づいて処理列を構築する部分までは従来通り前向き推論を用い、値に即した部分つまり定量的な判断を行なう部分に非単調推論の一実現形式である TMS を利用することとした。以下では、非単調推論と TMS について説明する。

4. TMS について

4.1. 非単調推論

人工知能の分野において、計算機上での知識表現あるいは推論のために一階述語論理が用いられている。一階述語論理の世界において知識は単調増加である。新たな知識が加わると全体の知識は増加するのが普通であり、一度真とされた述語ののちに偽となることはない($P \wedge \neg P$ は常に真とならない)。しかし、現実世界では「雨が降らない限り晴れている」(形式的に記述すれば P

unless $\neg P$) など、新たな知識が得られた時点で以前の知識が不成立となることも多く、知識量の変化は非単調である。

このような問題に対するアプローチのひとつとして、非単調推論の研究が行なわれている。TMS (Doyle, 1979) はそのうちの一実現形式である。

4.2. TMS の概念

以下では TMS の原論文に従い、各々の述語を「信念」という言葉で表す。

TMS では、個々の信念は *in*, *out* のいずれかの状態を取る。ある信念が *in* か *out* かは「状態 *X* は、他の状態 *Y* が *in* であり、*Z* が *out* の時、*in* である」などの依存関係により定まる。この依存関係はサポートリスト (Support List) と呼ばれるリストで表され、以下の二種類がある。

SL: (SL *in-list out-list*)

in-list 内の全ての信念が *in* であり、*out-list* 内の全ての信念が *out* の時、このリストが対応する信念は *in* である。

CP: (CP *result in-list out-list*)

in-list 内の全ての信念が *in* であり、*out-list* 内の全ての信念が *out* の場合には常に *result* が *in* となるとき、このリストが対応する信念は *in* である。

サポートリストが存在することを、対応する信念は正当化 (Justification) された、という。

サポートリストが空である場合には状態は変化しないため、常に真である信念 (前提: Premise) を定義することができる。また、*in-list* が空でなく *out-list* が空である SL サポートリストを持つ信念は「～ならば…である」という含意を表すと考えることもできる。

ある信念 *A* に対する SL サポートリストにおいて *out-list* が空でない場合、*out-list* に書かれている信念の状態によって *A* の正当化は変化する可能性がある。このような信念 *A* を「仮説」と定義する。TMS では他に、矛盾 (CONTRADICTION) という特殊な信念を導入している。

TMS のアルゴリズムを簡約すると、

1. 新たな信念の入力を受けて、他の信念の状態更新を行なう。
2. CONTRADICTION (以下 *C* と略) という特殊状態が *in* になっている場合、
 - (a) *C* の *in-list* から *C* を *in* としている「仮説」の冗長でない最小集合を計算し、これを *S* とおく。
 - (b) (CP *C S* ()) という正当化のもと、NG (NoGood) という状態を作成する。効率化のため、CP に関して完全な操作は行なわれず、その依存関係だけが考慮される。
 - (c) *S* の中から任意の仮説を一つ選び、その仮説の *out-list* の要素の一つを、

(SL (NG と *S* の残り) (選んだ *out-list* の要素の残り))

で正当化する。これにより、選択された仮説と *C* が *out* に、また上記で正当化された要素は *in* になる。

わかりやすく言えば、矛盾に至る根本となった仮説の一つを否定し、矛盾に至ったこと自体を否定の理由とすることになる。

3. *C* を *in* とする別な正当化があれば、再度 2 以下のプロセスを行なう。矛盾がなくなれば、終了。

なお、TMS 自体は問題解決器ではなく、あくまで述語間の相互関係を調べて仮説の真偽を推移

させ、無矛盾の状態を維持するだけである。問題解決を行うための機構は別に用意しなければならない。

本システムでは、形式的知識に基づいて作成された処理列の実行を、TMS に即して用意した前向き推論機構によって行なう。

5. TMS を用いた探索アルゴリズム

これまでに述べた問題点の解決を図るため、知識管理、探索面で TMS の導入を行ない改良を加えた探索アルゴリズムは以下の通りである。

1. ユーザからの観測データと解析目的の入力を基に中心処理を決める。

システムは観測データから得られる定性的情報と解析目的に基づき可能と思われる中心処理候補を探索する。

各候補の知識ベースには中心処理として用いられるための条件が書かれている。システムでは、観測データとそれらの条件が部分的にでも合致し、かつ明らかに反していない場合には、当該処理を中心処理の候補として選択する。さらに、条件のうち未定である属性のリストを作成し、あらかじめ属性名に割りつけられた重みに従った順序で並べ変える。並べ変え後のリストに従ってシステムは未定属性に対する質問を作成し、ユーザに入力を求める。

ユーザは Yes, No あるいは選択枝から一つ選ぶなどの方法でシステムからの質問に答える。回答から得られた情報により未定条件属性は減りシステムは更に候補を絞っていく。

質問の内容がわからない場合などは、ユーザは質問に対して「不明」と回答することができる。この場合、システムは「不明」とされた属性を検索対象から外し、プロセスを続行する。このプロセスは差分リストが空になるか未定条件属性の値が全て「不明」となるまで続けられる。

2. 観測データ及びユーザから入力された形式的知識に基づいて処理列を構築するための問題解決処理を行ない、観測データから中心処理までの処理列および中心処理から出力処理までの処理列をすべて求める。

個々の処理列候補について、「不明」などによらず多くの属性を条件として用いたものから順序づけが行なわれ、それぞれの処理列が TMS の仮説として管理される。

3. 2 で得られた作成処理列を順序に従って採択、実行する。

処理列における個々の処理において、処理前にデータの性質をチェックしなければならないものはテストを行なう。REX と同様、結果に関する閾値に抵触する場合には、変数変換、変数選択などの前処理が可能であればそちらを用いたのち、再度テストを行なう。それでもなお条件に抵触する場合には、矛盾状態を宣言する。

また、手法実行に際して引数が必要なものについては、まず知識ベースに記されているデフォルト値が仮説として定義され、それに基づいて処理が実行される。実行後、得られた結果に対する指標値の計算やテストを行ない検証を試み、失敗と判断された場合には、矛盾状態を宣言する。

矛盾状態となった場合には TMS アルゴリズムに従い、現在試行中の処理列全体と、矛盾に至るまでに用いられた仮の引数、仮定された属性などが集められ、それらの中から任意の 1 仮説を否定することで矛盾状態を解消しようとする。

その際、どの仮説を否定するか、また選択された仮説を否定するための正当化には仮説を成立させている *out-list* のどの要素を用いるかが問題となる。選択方法について TMS のアルゴリズムでは特に規定されていないが、本システムでは効率化のため、対象処理列のうち現在着目している手法と矛盾対象との距離(間にある処理数)や否定された要素自身の重みで判断している。また、本システムで扱われる仮説は、自身の成立が他の仮説に依っていないため、*out-list* の要素は 1 つであり、そのための選択機構は不要である。

6. システムの構成

本システムは推論処理部を Prolog で、データ解析作業は S 言語で、モジュール間のインタラクション及びユーザとのインタフェース部分は C 言語を用いて記述されている。

推論処理部で用いている論理型言語 Prolog では、一階述語論理相当の記述を必要とする。そのため、S 言語の dump コマンドで S 式 (S-expression) に変換した変数内容を UNIX 等で用いられている字句解析系 lex を用いて作成したトランスレータを用い、Prolog のファクト節に変換して用いている。個々のモジュールは独立に構築されており、S 言語に対する依存性は比較的少ない。同様のトランスレータを用意することで、他の解析処理系への対応も可能と思われる。

本システムのような仕様を考えた際には、解析処理列を構築する際に用いるための処理手法に関する静的知識の他、解析処理の途中における結果、仮引数などの動的な知識も管理しなければならない。両者は手続き上区別される場合もあるが、本来一貫した知識であり、可能であれば一元管理されることが望ましい。本システムではフレーム理論を参考とした知識管理モジュールを構築、利用している。これらに関する詳細は南ほか (1993) を参照されたい。

7. 実行例

ここでは、既知の解析例を再試行して結果を確認した例と、単一処理では実行できないデータへの本システムの適用例を示す。

7.1. 単一処理列の示唆・実行

塗料用のポリマーの開発に際し、ビニルモノマーに対して、溶解性、塗膜への強靱性、化学的安定性、顔料の分散性をそれぞれ与えるモノマーとの重量パーセント比、以上の共重合で得られたポリマーと無機顔料との親和性、ポリマーの重量平均分子量 P_1 、数平均分子量 P_2 、分子量分布の一般的指標となる P_1/P_2 、塗膜表面の光沢度 (2 つ)、塗料の粘度、塗膜中の無機顔料の均一な分散の度合いの 12 変量のデータ (吉澤 (編), 1993) に対する本システムでの実行例を以下に示す。なお、データはあらかじめ S 言語の変数として読み込んでおく必要がある。

まずシステムは、データの形状から個体-項目といったデータ行列であることを初期情報として持つ。ついで、この形状に沿う中心処理候補を知識ベースから収集し、それぞれの処理に関する特性情報の差分を作成したのち差分を補う情報をユーザに求めていく。

この例の場合にはまず解析目的が問われるので、ユーザは「指標値を得る」意味に近い simplification を選択する。システムは simplification を目的とする処理候補として知識ベースから回帰分析、主成分分析、数量化 III 類を得、それぞれを選択するための定性的情報の差分を作成

する。具体的には、回帰分析については説明変数と目的変数という形での外的基準が、数量化 III 類については入力データの尺度が名義尺度であることなどを抽出しているのが、実行例に示されている。

属性に関する重みづけに従って並べ替えられた順序に従い、ここでは次に外的基準の有無がユーザに問われる。各項目に対する事前情報やモデルは仮定していないため、ユーザはここで not exist を入力する。この結果、システムでは回帰分析を選択対象から除外する。

```
Welcome to our system

Please type in the name of data (in S)
sampTMS

name=sampTMS
type=data
size=44
item=12

What is your purpose?

1: simplification
2: scaling
3: estimation
4: classification
5: unknown

Select [1-5]: 1

SearchInfo:
  extcrit:[[=,exist]:[regression],[=,none]:[quantification3,pca]]
  scale:[[=,ns]:[quantification3],[>=,os]:[regression,pca]]

External criterion :
  Please select one. If you don't understand, please select ''unknown''

1: not exist
2: exist
3: unknown

Select [1-3]: 1

SearchInfo:
  scale:[[>=,os]:[pca],[=,ns]:[quantification3]]

Scale :
  Please select one. If you don't understand, please select ''unknown''

1: Interval/Ordinal Scale
2: Nominal Scale
3: unknown

Select [1-3]: 1

Purpose: simplification
CenterProcedureCandidateList: [pca]
QualitativeOutput: [pca],[item=pca:p,type=data,size=44,scale=is,extcrit=none]

CommandS: sampTMS1 <- pca(sampTMS,2)
```

図 3. 7.1 の解析支援実行過程（下線はユーザが入力した部分）

さらに観測尺度が間隔尺度であるかどうかが問われ、ユーザが間隔尺度を選択することにより数量化 III 類が候補から外れ、システムは中心処理候補として主成分分析を選択する。

観測値はそのまま中心処理にかけることが可能なため、観測値から中心処理に至る処理列は不要となり、直接主成分分析が行なわれる。その際、主成分の個数の初期値として仮説 $p=2$ を用いるが、出力結果によれば第 2 主成分までの累積寄与率は 0.4107 と低い。主成分分析に関する知識ベースに記述された事後条件として 0.8 以上が指定されている。従って、ここまでの実行過程に矛盾を生じることとなり、矛盾状態解消のために TMS プロセスが起動される。

実際には、主成分分析の実行を成立させていた仮説集合のうち、出力主成分数を指す仮説 $p=2$ を *out* とする。これにより累積寄与率に関する矛盾状態が解消されるとともに、 p に対して定められているルールに従い、新たに $p=3$ とされ再処理される。しかし、これも知識ベースに記述されている条件を満たさないため、最終的には第 4 主成分 ($p=4$) までを結果として得ることになる。

中心処理されたデータをどのように出力するかについても同様のアルゴリズムで対処できるが、出力方法はユーザが結果をどのように解釈するか、どのように用いるかに依存する部分が大きいが、これはすなわち領域知識の利用につながり、システムによる完全な支援は不可能であることを意味している。実際、階層的クラスタリングに対するデンドログラムなど、一部の出力結果以外は、形式的な知識だけから出力方法を限定することは難しい。そのため、本システムでは特に、ユーザが出力方法を指定することも認めている。

実行例では出力方法を指定しなかったため、得られた主成分の個数、つまり 4 変量に対する表示法として一般的な対散布図が選択された。

中心処理決定までの実行過程を図 3 に、最終的な出力結果である対散布図を図 4 に示す。ユー

図 4. 7.1 節に関する解析処理列の出力結果

ザは出力結果に対し Brushing などの機能も援用した上で、解釈を試みることになる。

7.2. 複数処理列の構築・実行

表1は44人に対して6種類の酒類に関する嗜好を調査した0-1データである。このデータを「分類」することを目的としてシステムを起動する。

システムはまず、「分類」というユーザからの解析目的指示を受けて、中心処理を判別分析、数量化II類、クラスタ分析に絞る。ついで、前例と同様に外的基準に関する質問をユーザに発し、これも外的基準を有しないため、クラスタ分析だけを候補として選択する。この時点で中心処理は定まるが、観測データが名義尺度であることは以後の検索のために取り込む必要があるため、システムでは尺度に関してもユーザに質問を行なう。

結局、中心処理はクラスタ分析となるものの、前例と異なり観測値を直接適用することができないため、システムでは問題解決機構を用いて、処理列を構築する。その結果、データ間の距離を算出した上で階層クラスタリングを行なう、という処理列が導出されることとなり、システム側で対応可能な処理列として、

1. 二値距離、階層クラスタリング、デンドログラム
2. 数量化III類、Euclid距離、階層クラスタリング、デンドログラム

の2処理列が得られる。

現在のシステムでは処理を極力S言語組み込みのものに依っているため、ここでのクラスタ分析は階層クラスタリングとなり、`hclust()`の引数として指定可能な、階層クラスタリングの手法名が仮説となる。なお、仮説の初期値は、S言語における`hclust()`のデフォルトである最長距離法である。

さて、1で提案された処理列の実行結果を図5(a)に示す。一応、数群に分かれてはいるものの、この結果に対してユーザ側では明確な意味づけを行ないにくい。そこでユーザがシステムに対して再試行を要求する。この処理列における仮引数、つまり矛盾により否定、再配置されうる仮説は階層クラスタリングの方法のみである。従って、仮説変更により、群平均法によるクラスタリングが行われるが、この結果もユーザからみて明確な結果とならず、ユーザが再試行を促し、

図5. 二値距離を用いたクラスタリング結果（左から、最長距離法、群平均法、最短距離法の順）

表 1. 酒類の嗜好に関するサンプルデータ

No.	日本酒 1	日本酒 2	焼酎	ワイン	ウイスキー	ビール
1	1	1	1	0	1	0
2	1	1	1	1	1	0
3	0	0	0	1	0	1
4	1	1	0	1	1	0
5	0	0	1	0	0	1
6	0	0	0	0	0	1
7	1	0	1	0	0	1
8	1	1	1	1	0	0
9	1	0	1	1	1	0
10	0	0	0	1	0	1
11	1	1	1	1	0	1
12	1	0	1	0	1	1
13	1	1	1	1	0	1
14	0	0	1	1	1	0
15	0	0	1	0	1	1
16	0	1	0	1	0	1
17	1	1	1	0	1	1
18	1	1	1	1	0	0
19	1	1	1	1	0	1
20	0	0	0	0	1	1
21	1	1	1	1	0	1
22	1	1	1	0	1	0
23	1	0	1	0	0	1
24	1	1	1	0	0	1
25	0	1	0	1	1	1
26	1	0	0	0	0	1
27	0	1	0	1	1	0
28	0	0	0	0	0	1
29	0	0	1	1	1	0
30	1	1	1	0	1	0
31	0	1	0	0	1	1
32	1	0	1	0	0	1
33	1	1	1	0	0	1
34	1	1	1	1	1	0
35	0	0	1	0	0	1
36	1	1	0	1	1	1
37	1	1	1	0	0	1
38	0	0	1	1	0	1
39	0	0	1	1	1	0
40	1	0	0	1	1	1
41	1	1	1	0	0	0
42	0	1	0	1	1	0
43	1	1	1	0	1	1
44	1	0	0	1	1	0

最短距離法での結果を得る。

図 5 (b), (c) にそれぞれ群平均法, 最短距離法による出力結果を合わせて示す。しかし, これらにおいても構成されたクラスタに関する意味づけがしにくいとユーザが判断した場合, 再び再試行を促すこととなる。

その結果, 今度は 1 の処理列全体が矛盾状態とされ, 新たに 2 の処理列を仮説として選択する。

この時の処理列と仮説に関する概念図を図 7 に示す。2 の処理列で実行される数量化 III 類において, 得られた固有値に対する定量的な判断が困難なため, 最小である 1 変量の出力が行なわれる。得られた変量について Euclid 距離(無論, 1 変量の場合には絶対値となる)を取ったのち, や

図 6. 数量化結果から 1 変量を用いた解析結果 (最長距離法)

図 7. 処理列の概念図 (楕円で囲まれているのが仮説)

はり仮決定された最長距離法による階層クラスタリングが行われる。つまり、処理列中の2つの処理において、それぞれ仮決定部分が存在することになる。

実行の結果得られたデンドログラムを図6に示す。しかし、図6では明確な特徴を読み取りにくく、ユーザがまた再試行指示を出す。

図8. 数量化結果から1変量を用いた解析結果（群均等法）

図9. 数量化結果から1変量を用いた解析結果（最短距離法）

本処理列のように、仮決定した引数などが複数存在する場合、従来作成したシステムでは推論機構の制約により、まず、最も最近行われた手法の中で調節を試みる。つまり、階層クラスタリングの手法を変更することで対処しようとする。実際に旧システムで実行すると、階層クラスタリングの手法をそれぞれ群均等法、最短距離法に変更して再試行を行なう。この結果得られたデンドログラムを図8、図9にそれぞれ示すが、図6と比べても、出力結果から顕著な特徴を見い出すことができない。この処理列において重要な意味を持っているのは数量化による出力変数数であり、クラスタリング手法を変更する前に、数量化 III 類の結果から得られる固有値を勘案して出力変数を2変量、3変量と増やすのが一般的な戦略と考えられる。しかし、既存のシステムでは順次後戻りを行なうため、既に示した通り、複数の階層クラスタリングの実行という、冗長な過程を強いられる。

本論文で説明している TMS を用いたシステムでは、矛盾解消を行うための仮説選択の際、階層クラスタリングの方法よりも数量化の変数数に高い優先順位を与えておくことで、数量化の再試行を優先させることが可能である。実際に本システムによって再試行を行なったところ、クラスタリング手法の変更ではなく、数量化による変数数を2として、以下の処理列が順次実行された。その結果得られたデンドログラムを図10に示す。

図6と比較すると、どちらの結果においてもデータ番号の6や28のようなビールのみを好むクラスは明確に現れているが、図10の4から44までの組に対してはビールを好まないグループ、3から20については国産の酒を好まないグループと考えることができるなど、これまで挙げた図よりも図10の方が得られたデンドログラムに対して明確な解釈を与えることができる。

なお、システムから提案されたどちらの処理も、形式的に判断すれば全く問題のない組み合わせであり、特に表1のようなデータを解析する場合、1に示した処理列が想定されるのが普通である。本例では、S言語に用意されていた距離計算法がデータに合わなかったと考えられ、他の距離計算法、例えば Hamming 距離を取る処理がシステムに実装されていれば、その段階で問題のな

図10. 数量化結果から2変量を用いた解析結果（最長距離法）

い結果を得ることができたと思われる。これは、0-1 データを連続量に変換する処理である数量化 III 類がたまたま次元縮小の効果を合わせ持っていたため得られた処理列が効果的に働いた例と考えることができる。

ユーザとしては現在保有するシステムで可能な、かつ統計的に問題のない結果を得ることが理想であり、その意味で本システムのような支援形態は、新たな処理を自動作成できないなどの限界はあるものの、支援として一定の役割は果たせると考えられる。

8. おわりに

本論文では、データ解析支援システムにおける知識処理、特に推論機構に関して従来の問題点を具体的に指摘した上で、望ましい推論機構に関する考察を踏まえて、TMS 利用による推論機構を導入した支援システムの提案とその構築例を示し、実行例により有効性を確認した。これにより、仮引数などを仮説とし、仮説間の優先順位を考慮することで、統計家が実際に行なう解析作業の順序に近い実行手順の支援が可能になったと思われる。

今後、以下のような課題が考えられる。まず、本システムは、システムの応答に対するユーザの疑問について説明を行なうような、説明の能力が不十分である。REX, RASS はそれぞれ説明機構を有しており、システムが選択したあるいは行なった処理内容についてユーザに理解を促すことができる。しかし、複数手法を対象とする本システムで十分な説明を提示するためには、両システムよりも更に一般的な知識の実装が必要と考えられる。限られた知識内容からどのようにして効果的な説明を行うかが、今後の課題のひとつである。

加えて、非単調ではあるが、TMS でも知識量の増加、それに伴う探索効率の低下は著しい。このような欠点に対するアプローチとして de Kleer (1986) による ATMS (Assumption-based TMS) が既に報告されており、今後に向けて利用を検討中である。

最後に、本システムのような手法選択支援では、個々の処理に関する知識に共通性を持たせる必要があること、また既存の解析システムに付加する形を取ることから、さまざまな制約が生じる。本論文で挙げた解析例においても、主成分分析において累積寄与率に関する判断を支援システム側で行なうために事前に主成分の個数を与える形を取っている。また、数量化 III 類についても 1 次元数量化から試行し、矛盾によって順に次元を上げることとしているが、固有値に関する適切な判断が可能であれば、試行錯誤による必要はない。また、階層クラスタリング手法と数量化の次元数との重みづけも統計家によって異なることが充分考えられる。このような問題は、知識ベースの語彙を増やし記述能力を上げることで若干の改良は可能であるものの、判断要素が複雑になればなるほど、知識ベースを構築する際に用いられる知識にも共通性が欠けていくことになるため、今後の改良方針とあわせて、十分な考察が必要である。

参 考 文 献

- Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988): *The New S Language*, Wadsworth & Book/Cole Advanced Books & Software, Pacific Grove (渋谷政昭, 柴田里程訳 (1991): S 言語 I, II, 共立出版)。
- Brent, E.E. (1988): *STATISTICAL NAVIGATOR: An Expert System to Assist in Selecting Appropriate Statistical Analysis*, Columbia, MO: The Idea Works, Inc.
- Doyle, J. (1979): A Truth Maintenance System, *Artificial Intelligence* 12, pp. 231-272.

- 古川康一, 溝口文雄 (編) (1988): 知識プログラミング, 共立出版.
- Gale, W.A (ed.) (1986): *Artificial Intelligence & Statistics*, Addison-Wesley.
- Hand, D.J. (1993): Measurement scales as metadata, in *Artificial Intelligence Frontiers in Statistics*, Hand, D.J. (eds), Chapman & Hall, pp. 54-64.
- de Kleer, J. (1986): An Assumption-based TMS, *Artificial Intelligence*, Vol. 28, pp. 127-162.
- Minami, H., Mizuta, M. and Sato, Y. (1993): A Knowledge Supporting System for Data Analysis, *Journal of the Japanese Society of Computational Statistics* Vol. 6, No. 1, pp. 85-97.
- 南 弘征, 水田正弘, 佐藤義治 (1993): データ解析支援システムにおける知識ベースの枠組とその実現, 計算機統計学 Vol. 6, No. 1, pp. 37-48.
- 中野純司, 山本由和, 岡田雅史 (1991): 知識ベース重回帰支援システム, 応用統計学 Vol. 20, No. 1, pp. 11-23.
- Oldford, R.W. and Peters, S.C. (1988): DINDE: Towards more sophisticated software environments for statistics, *SIAM J. Sci. Stat. Comput.*, No. 9, pp. 191-211.
- Streitberg, B. (1988): On the nonexistence of expert systems: Critical remarks on artificial intelligence in statistics (with discussions), *Statistical Software Newsletter* 14, No. 2, pp. 55-62.
- 吉澤 正, 芳賀敏郎 (編) (1992): 多変量解析事例集第1集, 日科技連.

(1993年10月8日受付 1994年7月24日最終修正)

著者連絡先: 〒047 小樽市緑3-5-21 小樽商科大学商学部社会情報学科 南 弘征