北海道大学工学部研究報告
第167号（平成6年）

Bulletin of the Faculty of Engineering
Hokkaido University, No. 167 (1994)

TMSを用いたデータ解析支援システム

南　弘征　水田　正弘　佐藤　義治

（平成5年8月31日受理）

A data analysis supporting system with TMS

Hiroyuki MINAMI, Masahiro MIZUTA and Yoshiharu SATO

(Received August 31, 1993)

Abstract

Most data analysis systems seem to be convenient for statisticians, but novices in statistics can not use them easily since it is hard to select appropriate procedures according to data and purpose.

We had already proposed the data analysis supporting system with the techniques of knowledge processing but there remains a problem about the efficiency of reasoning. Since the reasoning method depended on Prolog system completely, we processed some unnecessary steps.

TMS(Truth Maintenance System) is one of the assumption based reasoning methods. We offer a data analysis system with supporting function based on TMS and show an example of execution.

1．はじめに

近年，データ解析において計算機の存在は不可欠であり，S言語[1]，SASなど計算機上での優れた解析パッケージが数多く開発され，解析手法，データの表示法も解析作業を行なうためには量，質ともに充分なものがサポートされている。また，X-Windowなどの環境利用，マウスなどによる操作性の向上など，ユーザインタフェースも発展しており，システムの操作もより簡便に行なえるようになってきている。

しかし，システムの発展が直ちに解析作業全てを容易とするわけではない。パッケージによって可能となった手法が多ければ多いほど，データ解析，統計学に関する知識をさほど持たないユーザは特に，選択に戸惑い，結果として，誤用の可能性も多くなることが予想される。

このような背景から，データ解析そのものに関する知識を多く持たないユーザが，できるだけ正しい解析作業を行なえるよう支援するためのシステムの研究が1980年代から行なわれている。その代表的なアプローチとして，統計家が自身の経験などによって保持している解析戦略などを計算機上に実装，利用することを目的とした，データ解析エキスパートシステムの研究がある[2]。

これまでに著者らも，尺度を最小記述単位として，手法列の検索，構築を行なうシステムの提

情報図形科学講座

案を行なった[3]。しかし，推論方式としてPrologに依存した単純な前向き推論を用いたため，途中過程に関する情報を失うなど，処理が失敗した際のフィードバックが得にくい，という問題点があった。

以上のような背景に基づき，本論文では，TMS(Truth Maintenance System, Reasoning Maintenance System)[4]を用い，データ解析における手法選択，解析作業を支援することを目的とした，汎用的データ解析支援システムの提案とその構築例について述べる。

2．データ解析とエキスパートシステムについて

2．1 従来の研究例

一般にデータ解析と知識処理手法との融合を考える際，どちらを主とするかにより2つの立場がある。知識処理にデータ解析手法を適用した研究としては，意思決定などの問題解決に対するベイズ確率やShafer-Dempster理論の応用がある。本論文などで取っているアプローチは逆に，知識処理の手法を解析作業の過程に導入し，データ解析作業をより容易にしようという応用である。

このような応用例として報告されているものの多くは，特定の手法に対して，より細かな処理を行なう目的を想定している。例えば，データ解析でのエキスパートシステムとして先駆けといえるREX[2]は，CAI(Computer Assisted Instruction)的な狙いのもとに，システム内部で知識として持っている回帰分析の戦略に基づいた種々の指標値を計算し，ユーザの解析作業を支援するシステムである。同様に，回帰分析に関して，オブジェクト指向の概念を採り入れ，回帰分析の各過程を細分し，生じる個々の中間データを全て階層化してクラス分けを行ない，解析作業そのものはクラスとオブジェクトのメソッドとして取り込んだシステムとして，RASS[5]がある。

一方，ユーザの解析目的に合わせ的確な手法を選択する類のものは少ないが，そのうちの1つにStatistical Navigator[6]がある。Statistical Navigator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，適切な手法を選択，実行する。しかし，手法の組み合わせまでは考慮されておらず，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。他に，やはり回帰分析用として発表されたSmalltalk環境で構築されているDINDE[7]などが他の手法への適用，また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。

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[8]が，REX, Statistical Navigatorなどはそれぞれ実用に供されており，他分野におけるエキスパートシステム同様，解析作業の支援に十分貢献すると思われる。また，このような研究例のほとんどが回帰分析を対象としているが，回帰分析が知識処理の応用対象として特に適しているわけではなく，回帰分析がデータ解析において基本的な手法であるためと思われる。回帰分析に限らず，他の代表的解析法にも知識処理のアプローチを用いることは可能であり，更にそれらをまとめ，選択するような枠組も可能であると考えられる。

本報告では，適切な解析処理の選択を統合利用の一形態として捉えることとし，以下では，知識処理による支援を目的とした，データ解析過程の定式化を考える。

2．2 問題の定式化

計算機による知識処理を応用するという前提に立ち，データ解析作業を大きく3段階に分けて考える。

1．計画：解析対象を決め，データを集める。
2．解析：集めたデータと解析目的に応じて適切な手法を選び，解析を行なう。
3．解釈：得られた解析結果から，データに関する結論を導出する。

データ解析において，そのデータがいかなる対象から生じたものであるか，解析結果に対してどう意味づけを行なうのか，つまり，「計画」，「解釈」においては，データ解析の知識に加え，領域固有の知識(domain knowledge)が重要な意味を持つ。しかし，特定領域に関する知識のみを多く有するシステムは，データ解析作業という側面からみると汎用性に乏しく，特定領域のみを対象とするエキスパートシステムと捉える方が自然である。汎用性を重視した場合，全ての領域知識を実装することが難しい現状では，できるだけ領域知識に依存しない部分を対象に支援システムを構築せざるを得ない。

領域固有の知識は「解析」でも用いられるが，「計画」「解釈」と比べた場合，データ自体を扱う実際の解析作業において領域知識の利用頻度は少ないと考えられる。従って，できるだけ汎用性を考慮したシステムを構築する，という主旨からは，2の「解析」過程を知識処理対象とすることになる。

更に「解析」自体も，原データと解析目的を入力し，合致した手法を選択したのち，実行，結果に応じた値の加工や手法の再検討を行なう，などの数段階に細分して考えることができる。この時，原データと解析目的入力，及びそれに基づく探索までは，検索対象とする情報が固定されている以上，決定的に行なうことが可能である。しかし一般に，探索によって選択された手法は唯一ではなく，結果に応じた試行錯誤が必要な場合も多い。また，個々の手法に応じた細かな副処理が必要な場合も考えられる。従って，以降の部分は非決定的に実行されなければならない。

また，データの具体的な値に言及することなく，データの構造，性質などにより，適用可能な処理をある程度形式的に選択できる部分もあるが，例えば回帰分析における多重共線性など，具体的なデータ値から得られる各種統計量によって適切な処理を選択しなければならない場合もある。本システムでは，手法選択の段階まではデータの定性的な情報のみで探索を行ない，加工，再検討の段階で定量的判断を含めた具体的な処理を行なうことにした。

以上から，本論文で取り扱う，汎用性を持ったデータ解析エキスパートシステムは，解析過程の手法選択を主眼に，手法の組み合わせを提案し，実行，再検討の作業を支援するシステムとする。

2．3　推論方式について

解析作業開始時に原データと解析目的があり，途中過程でデータを加工するための手法を選び，最終的に適切な出力法によって結果を得る，と考えれば，データを状態，特に原データを初期状態，解析手法を操作子，特に出力手法はデータを消失させる操作子とし，目標状態をデータが空である状態とすることで，探索過程は，人工知能の分野における問題解決のプロセスに帰着することができる。

問題解決において，推論方法や知識の記述方法はいずれも基本的な事柄である。推論方法は2つに大別され，初期状態が与えられ，それを変化させていくことで終了状態への到達を試みる前向き推論と，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がある。後向き推論は，最終状態から探索を開始するため，最終状態を実現する初期状態の数が少なければ，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，データ解析作業での問題解決には，後向き推論の適用は馴染まない。理由は，定量的な情報が主である解析作業において，実際に処理を実行することなく，つまり具体的な出力値を持たないまま，目標状態からの探索を行なうことが極めて困難なためである。従って，本研究のような目的では，前向き推論を選択することが妥当と考えられる。ところが，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。

これらの問題に対する枠組として，今回はTMSを利用することとした。

3．TMSについて

3．1　非単調推論

一般に，計算機上に実装される知識は単調増加である。つまり，新たな知識が加わることにより，全体の知識は増加するのが普通である。計算機での知識処理によく用いられる述語論理もやはり単調増加であり，ある述語が加わることにより，述語の真偽値が変化することはありえない。しかし，現実世界の知識において，新たな知識が成立した時点で以前の知識が不成立となることは多く，知識量の変化は非単調である。

このような問題に対して，非単調推論の研究が行なわれている。TMS[4]はそのうちの1つの実現形式である。

3．2　TMSの概念

以下ではTMSの原論文に従い，仮説とされる述語を「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，他の信念の状態と関係し，例えば，「状態Aは，他の状態Bがinであり，Cがoutの時，inである」等の依存関係を有する。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2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となる時，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)を持つ，という。

サポートリストが空である場合，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も考えられる。

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，この信念の正当化は変化する可能性がある。このような信念を「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。

なお，TMS自体は推論動作そのものは行なわず，あくまで述語間の相互関係を調べ，真偽を推移，維持するだけである。推論機構は別に用意しなければならない。

3.3 アルゴリズム

TMS の具体的動作原理は以下の通りである。

1．新たな信念の入力を受け，それに関する正当化を計算，他の状態の更新を行なう。

2．CONTRADICTION（以下 C と略）という特殊状態が in とされていない場合はここで終了。

3．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する。例えば，A，B がともに C の in-list にあり，A が B の in-list にあれば，A が最小である。以下この集合を S と置く。

4．(CP C S ())という正当化のもと，NG(No Good)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。

5．S の中から任意の要素 A_i ($i=1,2,\cdots,n$) を選び，A_i の out-list の要素を D_j ($j=1,2,\cdots m$) とし，そのうちの任意の D_k を，

(SL (NG A_1 A_2 $\cdots$ A_{i-1} A_{i+1} $\cdots$ A_n) (D_1 D_2 $\cdots$ D_{k-1} $D_{k+1}\cdots D_m$))で正当化する。これにより，

- A_i の out-list のうちの 1 つが in となるので，A_i は in ではなくなり，out
- A_i が in でなくなるため，C は out
- NG はそのまま in
- D_j は，C が矛盾であることから導かれる仮説であり，in-list，out-list ともそれぞれ in，out にあることは明確なので，in

となることがわかる。わかりやすく言えば，矛盾に至る根本となった仮説の 1 つを否定し，否定の理由づけとして，矛盾に至ったことを用いるのである。

6．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。

4．TMS を用いた探索アルゴリズム

4.1 前向き推論によるアルゴリズムと問題点

これまで，Prolog 処理系を用いた前向き推論による処理列構築を試みてきた[3]が，以下に挙げるような問題点があった。

まず，Prolog に依存した縦型探索による前向き推論の場合，探索に失敗し目標の再充足を行なう際，それまでの履歴を捨ててしまう，という特性がある。データ解析において，途中過程で生成される中間データは有用なことが多く，解析作業中に再利用する可能性が高いため，作業履歴，中間データは保存されるのが望ましい。しかし，Prolog に依存したままこれらを記述，利用することは難しく，従って，探索機構自体に別なアプローチが必要である。

また，探索自体が前向き推論の形態をなしていたため，例えば対数変換など，必要に応じて(demand-driven)起動される類の処理は，処理を 1 度失敗させることで再検索，再起動させていた。しかし，失敗による再充足の形では，前向き探索の特性上，失敗に至った原因を利用することができず，作業過程として自然な動作とは言えない。データ解析作業において，試行錯誤は極めて自然なアプローチであることを考え合わせるならば，状況依存型の再探索方法が有効である。

4.2 TMS を用いたアルゴリズム

以上の問題点の解決を図るため，知識管理，探索面で TMS の導入を行なう。TMS 導入による改良されたアルゴリズムは以下の通り。

1．ユーザからの原データと解析目的の入力を基に，中心処理を決める。

システムは原データから定性的情報を得，情報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，中心処理とするための条件が書かれており，システムは，属性名をキーとして，未定条件属性の差分リストを作成，あらかじめ属性名に割りつけられた重み順に並べ変える。並べ変え後のリストから，システムは質問を作成，提示し，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1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情報から外し，候補限定プロセスを継続する。

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．原データから中心処理までの処理列，中心処理から出力処理までの処理列を全解収集する。

個々の処理列候補は，より多くの情報により得られたものから順序づけが行なわれ，それぞれ仮説として管理される。

仮説概念の導入により，従来は不可能だった，1度失敗とされた処理列を，部分的に再利用できる可能性が生まれる。

3．処理を実行する。

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い，REXと同様，結果に関する閾値に抵触する場合は，この手法適用を矛盾とする。何らかの前処理を行なうかこの手法の適用を中断するかが矛盾解消の方法として考えられるが，この選択自体はTMSに依存する。

手法実行に際して引数が必要なものは，まずデフォルト値が仮説として定義され，実行後，得られた結果に対する指標値の計算，テストが行なわれる。テスト結果により失敗と判断された場合，デフォルト値変更，この手法の適用中断が考えられるが，この解決もTMSに依存する。

矛盾解消時，TMSの枠内で制御可能なものは，解消のために否定される仮説と，その仮説を成立させているout-listの要素である。どの要素を選ぶかは，現在実行している，あるいは実行し終えた手法，対象処理列中の，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)，否定された要素自身の重みなどで判断している。

5．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Prologは，述語論理に端を発する論理型言語であり，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，構文解析系lexを用いて作成したトランスレータを用いて，Prologのファクト節に変換して読み込んでいる。

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。

6．知識管理モジュール

処理手法に関する静的知識の他に，実際のデータが有する情報などの動的知識も管理しなければならない。本システムではフレーム理論を基に，知識管理モジュールを構築している[9]。

現在用いている静的知識ベースの例を図1に示す。purposeCondition:には中心処理として適用可能であるための条件が書かれている。他に，処理を直接適用するための条件や，処理実行のために必要なパラメータに関して，パラメータ名，初期値，増減方法，条件を併せて記述するようになっている。これらはいずれもPrologの述語に変換されて用いられる。

6台の車の「走り感」に関して，17項目について5点法で行なわれた評価値[10]に関する解析例を考える。車の差異を考えずに，「走り感」に対する指標を得たいと考えた場合の，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。

```
methodname: metricMDS
purpose: scaling
purposeCondition: [dist, dim > 1]
condition: [type=sqrt, symmetric, scale=is]
parameter: p
p->>default: 2
p->>restrict: [p =< dim]
p->>apply: increment(dim)
deleteAttribute: [type, dist, dim, symmetric]
addAttribute: [type=data, item=p]
callS: metricMDS(data, p)
```

図1　知識ベースの例：多次元尺度構成法

7．実　行　例

まず，データの形状から，正方行列ではないことを判断し，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。次いで，知識ベースから，この形状に沿う中心処理候補を収集し，それぞれの処理に関する特性情報の差分を作成し，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，「指標値を得る」意味に近いindexを選択する。次いで外的基準の有無が問われる。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，All sameを入力する。ここまでの入力で，システムは中心処理として主成分分析を候補としたが，観測尺度が間隔尺度であるかどうかの確認を，更に行なう。

以上の過程を経て，主成分分析が中心処理として選択された。観測値はそのまま中心処理にかけることが可能なため，直接主成分分析が行なわれる。その際，主成分の初期値として2が指定されるが，第2主成分までの累積寄与率は0.6065であり，主成分分析に関する知識ベースに記述された事後条件に反するため，矛盾を生じる。

ここで，TMSプロセスが起動される。TMSでは，主成分分析の実行を成立させている集合のうち，もっとも最後にin-listに加えられた仮説p=2をoutとする。これを受けて，pに対して定められているルールに従って，パラメータが新たにp=3とされ，再度主成分分析が行なわれるが，これも条件を満たさず，最終的には第6主成分までを結果として得る。

出力方法には，6変量に対する表示法として対散布図が選択される。実行過程の一部を図2に，対散布図を図3に示す。

```
Welcome to our system

Please type in the name of data (in S) tmssample
DEBUG: tmssample
CommandS: dump('tmssample', '/tmp/15119-tmssample')

What is your purpose?
1: index
2: classification

Select [1-2]: 1

scale is ' Interval Scale ' ? [y/n]? y
extcrit : Please select one.
          If you don''t understand, please select ''unknown''

1: The variables is '1 vs. other'
2: none
3: unknown

Select [1-3]: 2

Purpose: index
CenterProcedureCandidateList: [pca]

CommandS: tmssample1 <- pca(tmssample,2)
```

図 2　実行過程

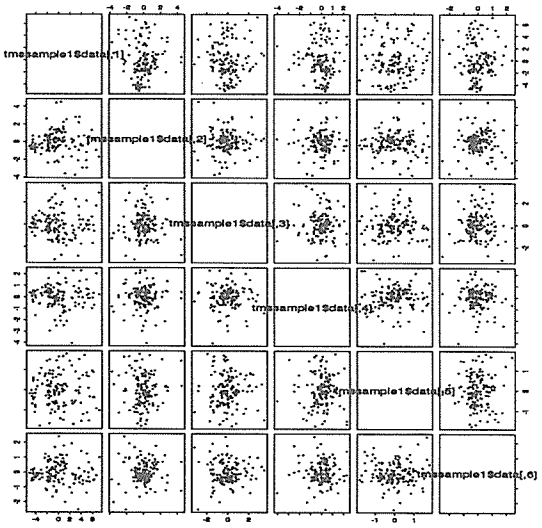

図 3　出力結果

8. おわりに

本論文では,データ解析システムに対してTMSを応用した支援システムの提案とその構築例を示した。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いるものの，回帰分析に重点を置いた知識構成となっており，説明できる内容に制約が生じている。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。説明機構の充実と合わせ，今後の課題の1つである。

また，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ATMS(An Assumption-based TMS)[11]が既に報告されている。ATMSは個々の述語の妥当性を検証することに力点がおかれており,現在のin状態を得るのが枠組上難しい。そのため今回は，具体的な処理列候補が単に得られるTMSを用いたが，今後に向けてATMSの利用を検討中である。

更に，知識の妥当性の検証，また，利用者を募った実験などを通じて，その有効性を評価するなどの実践的研究も行なう予定である。

参考文献

1) Becker, R. A., Chambers, J. M. & Wilks, A. R(1988) : The New S Language, Wadsworth & Brook/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).
2) Gale, W. A(ed.)(1986) : Artificial Intelligence & Statistics, Addison-Wesley.
3) 南　弘征，水田正弘(1992)：第60回日本統計学会講演報告集，pp. 245-247.
4) Doyle, J.(1979): Artificial Intelligence, Vol.12, pp.231-272.
5) 中野純司，山本由和，岡田雅史(1991)：応用統計学 Vol.20, No.1, pp.11-23.
6) Hand, D. J. (ed.)(1993) : Artificial Intelligence Frontiers in Statistics, Chapman & Hall.
7) Oldford, R. W and Peters, S. C(1988) : SIAM J. Sci. Stat. Comput., No. 9, pp. 191-211.
8) Streitberg, B. (1988) : Statistical Software Newsletter, Vol. 14, No. 2, pp. 55-62.
9) 南　弘征，水田正弘，佐藤義治(1993)：第61回日本統計学会講演報告集，pp. 138-140.
10) 吉澤正，芳賀敏郎（編）(1992)：多変量解析事例集第1集，日科技連.
11) de Kleer, J. (1986) : Artificial Intelligence, Vol. 28, pp. 127-162.