データマイニングにおける INDSCALの活用プロセス

奥　喜正

1．はじめに

IT時代ではデータを大量に容易に集計することが可能になってきたので，蓄積された大量データを効率的に圧縮して視覚化する新たなるデータマイニング技術の開発が要請されている(柴田，2001)。よって，従来よりもデータマイニングメソッドを利用する頻度が増してきた(福田・森本・徳山，2001)。それゆえ，統計的データ解析もデータマイニングへの新たなる貢献が求められてきているが，このような状況に貢献しているデータ解析メソッド(Computational Methods for Data Mining)には機械学習(Machine Learning)などから発展した，非確率構造モデル(Nonparametric Model)のクラスター分析，決定木(Tree Models)，ニューラルネットなどが，現時点では挙げられる(Giudici, 2003)。しかしながら，同様に非確率構造モデルに属する多次元尺度法(Multi Dimensional Scaling：MDS)は，データマイニングの実行時に利用される頻度が今までのところ少なかったように窺える(Kruskal, 1964)。

そこで，本稿では多次元尺度法のなかでも立体データの２相３元データ解析メソッドのINDSCAL(INdividual Difference SCALing：個人差多次元尺度法)を数年間の日本のプロ野球成績データに適用して(Carroll &, Chang, 1970)，３元多次元尺度法(３元MDS)のINDSCALをデータマイニングに有効活用する方法を検討する。INDSCALの応用例には，新製品の購入数予測に利用されたものがある(岡太・宮内，1996)。本稿では，３元データ解析でデータの経時的変動を適確に捉えることが可能なことや，データ圧縮過程で重要課題のデータ濃縮(Data Compression)をINDSCALがうまく遂行する様子と有効活用を示す。さらに，データマイニングの特徴を鑑みると，統計的データ解析メソッドの利用形態に多少の工夫を施してから，それらメソッドを利用する必要性がある。福田ら(2001)は，データマイニングでクラスター分析を使用する場合には，留意点として例外的なデータや不要属性の排除を指摘した。本稿では実際のデータ解析プロセスを通じて３元MDS利用時の留意点を検討し，データマイニングへのINDSCALの有用な利用形

態を模索する。

2．多次元尺度法とINDSCAL

最初に，多次元尺度法をクルスカルの方法(Kruskal, 1964)を例に説明する。多次元尺度法(MDS)とは，対象間の心理的距離など非類似性データから，潜在する少数の次元や背後にある構造を明らかにしたいときに利用される多変量データ解析手法の一つである。対象i, j, k, lにおける非類似性データδ_{ij}，δ_{kl}が得られたとすると，それらをユーザが視覚的に認識できる距離量d_{ij}，d_{kl}にできるだけその大小関係（順序関係）が保存されるように単調変換する。距離量とは三角不等式

$$d_{ik} + \leq d_{ij} + d_{ik} \quad \text{for } \forall\ i, j, k,$$

を満足する非負の量である。そして

$$\delta_{ij} < \delta_{kl} \quad \Rightarrow \quad d_{ij} \leq d_{kl} \quad \text{for } \forall\ i, j, k, l \qquad (1)$$

のように「クルスカルの単調回帰」を利用して，非類似性データ量を距離量に変換する(Kruskal, 1965)。各々の対象を点として表現する空間を布置と呼び，布置の次元をTとすると対象i, jの次元tにおける座標をx_{it}，x_{jt}のように求められれば

$$\delta_{ij} \overset{m}{=} \hat{d}_{ij} \approx d_{ij} = \sqrt{\sum_{t=1}^{T} (x_{it} - x_{jt})^2}$$

が成立することになる。ここで，記号$\overset{m}{=}$は単調増加関係を示し，d_{ij}は条件式（1）を満足する擬似距離量で，得られた布置のd_{ij}から計算される量でディスパリティとも呼ばれる。ところで，ある次元T(通常，T= 2，3)で完全に式（1）を満足するような布置を得ることは一般には困難である。そこで，与えられた非類似性データδ_{ij}の順序関係と，求められた布置から計算されるd_{ij}の順序関係がどの程度，一致しているかの評価基準にストレスSと呼ばれる，データに対するモデル適合度の評価尺度をMDSでは導入する。

$$S = \sqrt{\frac{\sum\sum_{i<j} (d_{ij} - \hat{d}_{ij})^2}{\sum\sum_{i<j} d_{ij}^2}}$$

順序関係が完全に保存されていれば，すなわち式（1）が成立すればS=0となり，ストレスSの値が0.2未満であることが最悪でも要求される。そうでない場合には，該当する次元ではデータ行列［δ_{ij}］に十分に適合する布置は得られないとみなして，通常は解の次元を一次元上げて，改めて布置を求める。

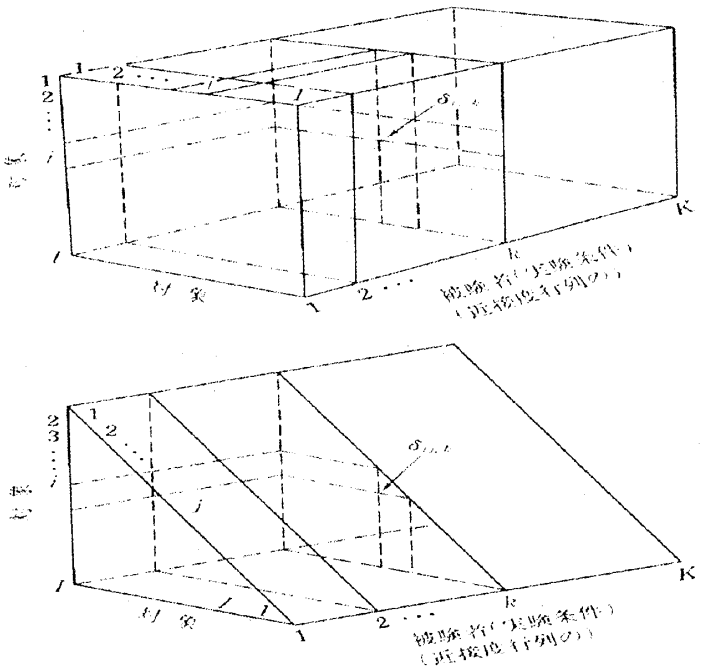

図１．３元データ（岡太・今泉，1990）

つぎに，３元MDSの代表モデル，INDSCAL(INdividual Difference SCALing：個人差多次元尺度法)について説明する。INDSCALは２相３元データを解析して(図１)，全体の被験者に共通する対象を布置する「共通対象空間Ｘ」と，被験者間の違い，すなわち，個人差を表示する「被験者空間Ｗ」を同時に出力するモデルである(図２)。被験者間の認知の相違を，布置次元に与える重みづけを被験者ごとに変えることで表現するモデルである。被験者ｋの対象ｉとｊの距離$d_{ij,k}$は

$$d_{ij,k} \approx \left[\sum_{t=1}^{T} w_{kt}(x_{it} - x_{jt})^2 \right]^{1/2} \qquad (2)$$

と定式化する。w_{kt}は被験者ｋ(ｋ＝１,･･･,Ｎ)が次元ｔ(ｔ＝１,･･･,Ｔ)を重要視する程度を示す重み係数である。x_{it}とx_{jt}は共通対象空間における，第ｔ次元の対象ｉと対象ｊの座標である。すなわち，x_{it}とは被験者全体に共通する共通対象空間での対象ｉのｔ座標であるのに対して，特定の個人ｋから対象空間を眺めた場合，対象ｉの布置は$\sqrt{w_{kt}}\ x_{it}$となって，被験者ｋの認知空間では座標軸ｔが$\sqrt{w_{kt}}$だけINDSCALでは伸縮されることになる。INDSCALは，被験者ごとに次元ｔで，$\sqrt{w_{kt}}$だけｔ座標軸を引き伸ばしたり縮めたりすることで被験者間の認知個人差を表現するモデルである(図２)。ここにINDSCALの３元データ解析モデルの使い易さがあると同時に，個人差を単に軸の伸縮だけで説明するという当該モデルの持つ限界も存在する。

ところで，式（２）の形では$d_{ij,k}$から，w_{kt}，x_{it}，x_{jt}の値を得ることはできない。

そこで，距離$d_{ij,k}$を，対象 i と対象 j を表すベクトルの内積$b_{ij,k}$にダブルセンタリング法(double centering)を用いて変換すると

$$b_{ij,k} = \sum_{t=1}^{T} w_{kt}\, x_{it}\, x_{jt} + e_{ijk}$$

と式（2）は書ける。ここで，e_{ijk}は誤差項である。この表現式にすると，$b_{ij,k}$ が与えられたときTの値を固定すると，$b_{ij,k}$ に最小２乗的にあてはまりのよいw_{kt}, x_{it}, x_{it}の値が正準分解(Canonical Decomposition Analysis)を利用して得られることが知られている。つまり

$$\sum_{i}\sum_{j}\sum_{k}\left(b_{ij,k} - \sum_{t=1}^{T} w_{kt}\, x_{it}\, x_{jt}\right)^2 \rightarrow \min$$

が成立するようにNILS(Nonlinear Iterative Least Squares)を使用して計算できる。このようにINDSCAL はモデル構築過程で内積を利用する。INDSCALの場合には，データに対するモデル適合度の尺度，ストレスSは

$$S = \frac{1}{m}\sum_{k}\left(\frac{\sum_{i}\sum_{j}[(\hat{d}_{ij,k})^2 - (d_{ij,k})^2]}{\sum_{i}\sum_{j}(\hat{d}_{ij,k})^4}\right)^{1/2}$$

のように定義される。

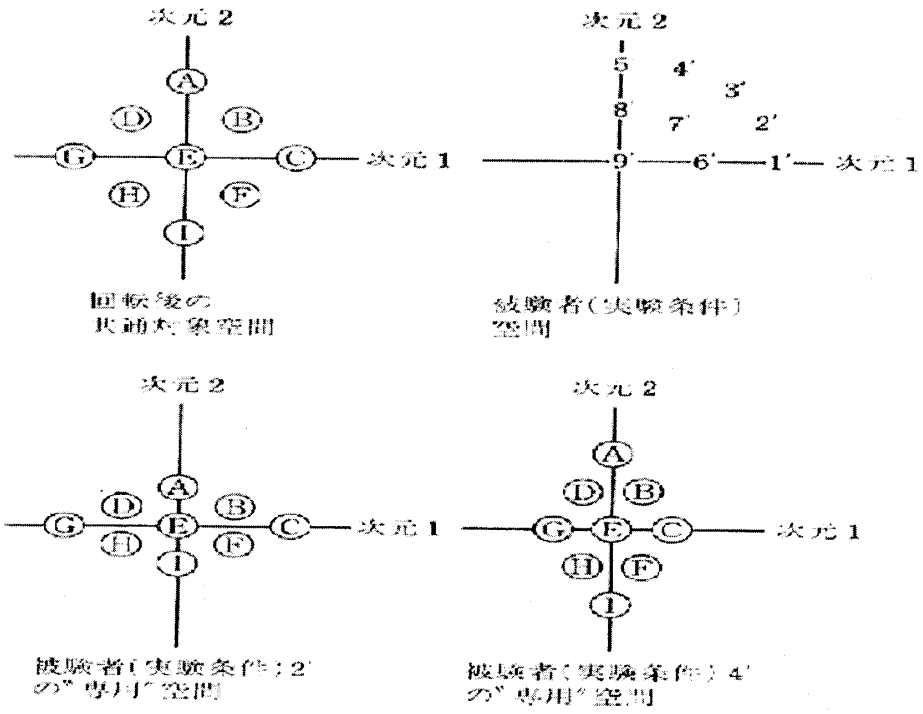

図２．INDSCALによる共通対象空間Xと被験者空間W（岡太・今泉，1990）

3．分析方法

1990年度から2001年度までの12年間にわたるプロ野球チーム12球団の成績データ，すなわち，６個の対象である成績，勝率，打率，得点，失点，本塁打数，防御率に関するデータをINDSCALで解析して対象６個の布置と年度格差を示す重みを得る。

一年間の成績データでは，12ｘ６型の２相２元データ行列Z_i（ｉ＝１，･･･，12)が形成される。対象間の非類似性はユークリッド距離を採用して計算すると，６個の対象間の${}_6C_2$個の非類似性データが計算されて単相２元データの６ｘ６型非類似性行列Δを生成する。12年間にわたってデータ集計したので12個の非類似性行列 Δ_1，Δ_2，･･･，Δ_{12}が作成される。さらに，これら12個の非類似性行列Δ_k（ｋ＝１，…，12)を縦に並べると，２相３元データの72ｘ６型の非類似性行列$D=[d_{ij,k}]$が形成される。

この非類似性行列ＤをINDSCALで解析して，２次元の共通対象空間Ｘと被験者空間Ｗ，そして，モデル適合度の指標であるストレス値を得る。今回のデータ解析では，布置の次元は一様に２に固定して，データ圧縮過程における情報損失に留意する。

4．結果

ストレス値が0.2を超える1992，1995，1996年度のデータは次元数２でINDSCALモデルで分析するのは不適切であると見做して３元データ解析対象から外し，残り９年間のデータのみで改めてINDSCALによる解析を実行した。このようなデータマイニング的な対処によって(福田他，2001)，平均ストレス値を0.166から0.119にまで低下させることに成功した。一年間のデータだけでは安定した対象布置は得られないことは実証されていて(奥・前鶴，2002)，９年間の成績データから計算された非類似性行列Ｄに基づいて，対象に関する安定した布置を得たが，特にストレス値が大きいデータは当面の解析対象から除いて，これらデータは別途，他のメソッドで分析する必要がある。

INDSCALによる共通対象布置X_1(12年間)，X_2(９年間)，被験者空間布置Ｗを図３，図４および図５に示す。INDSCALの被験者空間Ｗという出力は，プロ野球成績データの時系列的変化を捉えたものと理解できて(図５)，３元データ圧縮過程でのデータ濃縮という課題をINDSCALが首尾よく遂行していることが窺えた。12年間のデータによる布置X_1(図３)とモデル適合度が良い９年間データの布置X_2(図４)が，それ程異なっていないことが確認でき，３年間のデータ排除は総括的なデータ解析に大きな影響を与えなかったものといえる。

表１．12年間のINDSCAL解析結果における各年度の布置のストレス値と平均ストレス値.

年度	Stress	RSQ	年度	Stress	RSQ
1990	.116	.920	1991	.099	.940
1992	**.279**	.553	1993	.141	.877
1994	.172	.838	1995	**.227**	.694
1996	**.246**	.660	1997	.137	.885
1998	.097	.957	1999	.115	.922
2000	.138	.904	2001	.093	.960

Averaged (rms) over matrices

StrPPess = .16614

表２．９年間のINDSCAL解析結果における各年度の布置のストレス値と平均ストレス値.

年度	Stress	RSQ	年度	Stress	RSQ
1990	.100	.943	1991	.094	.957
1993	.110	.929	1994	.188	.818
1997	.130	.912	1998	.105	.953
1999	.087	.958	2000	.124	.930
2001	.099	.957			

Averaged (rms) over matrices

Stress = .11871

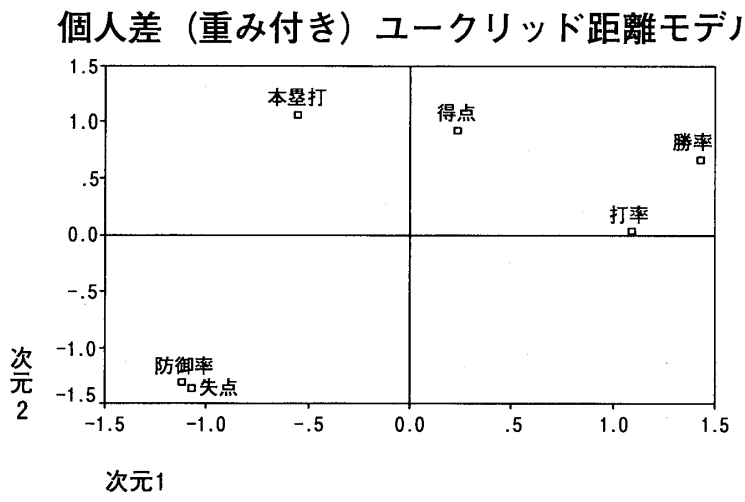

図３．INDSCALによる12年間の共通対象空間の２次元布置.

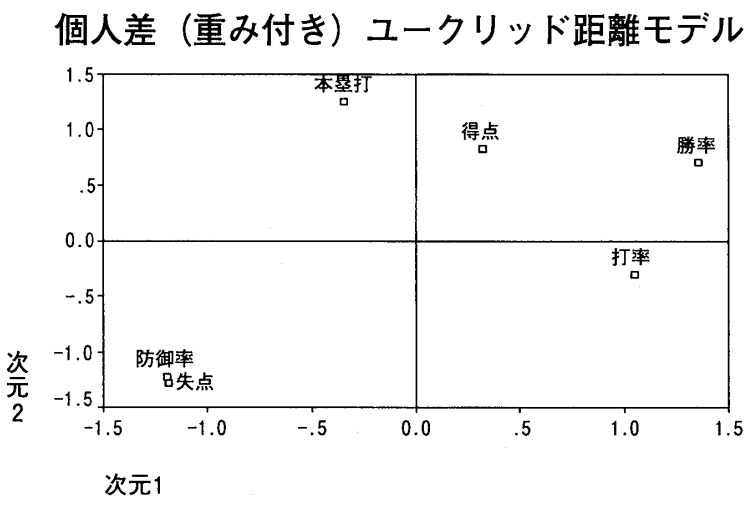

図４．INDSCALによる９年間の共通対象空間の２次元布置.

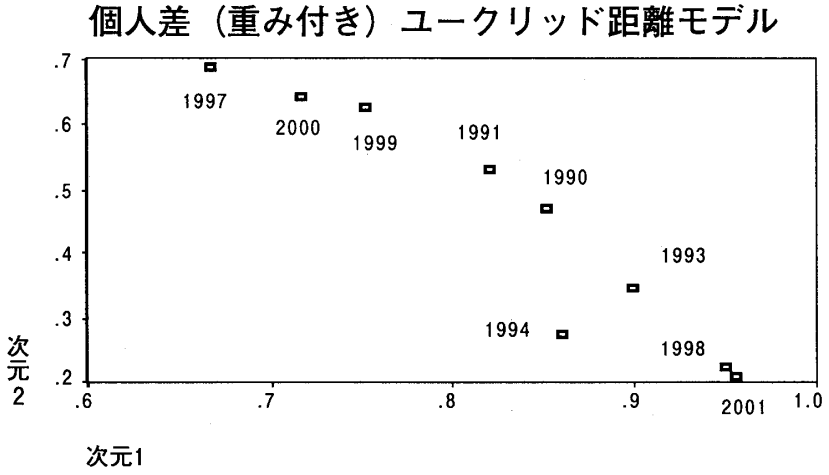

図５．INDSCALによる９年間の被験者空間 W.

５．考察

２相３元データをクルスカルの方法で解析した場合に比べてINDSCALを利用した場合のほうが，被験者空間Wが出力されるだけデータ圧縮過程における情報損失を減らすことができる(奥，2002)。すなわち，２相３元データのもつ経時的変動をINDSCALは被験者空間Wで捉えることができて，INDSCALが時系列的な３元データを解析するのに適切な方法であることが窺える。データ圧縮過程でデータサイエンスが目指すデータ濃縮，データ集約という作業に(柴田，2001)，INDSCALが参画して有用な情報を形成している様子が理解できた。すなわち，９年間のデータから成る２相３元データを

INDSCALのよって解析した結果の共通対象空間Xは，6個の対象の安定した布置となっており，同時に，被験者空間出力Wは年毎の布置変動を示すものと解釈できる。

INDSCALによるモデリングは年毎の布置変動を軸の重みの変化として捉えて，プロ野球成績の年度毎の特徴とみなして個人差，このケースでは年度差を被験者空間Wで捉えるモデル構築の立場を採っている。これに対して成績の年度差を偶然的変動とみなす立場では，9年間のデータを全体として総括してから，クルスカルの方法を一回利用したデータ解析を実行すればよいことになるが，この場合には成績の年度差は考慮しないから，データの経時的変動を無視されることを覚悟しなければならない。集計されたデータが経時的性質を伴う場合には，経時変化を無視せずに時系列的変化を捉えることが可能なINDSCALのような3元データ解析モデルを使用してデータ圧縮時の情報損失を少なくすべきであろう。

被験者ごとに次元tにおいて，$\sqrt{w_{kt}}$だけt座標軸を引き伸ばしたり縮めたりすることで被験者の認知空間個人差をINDSCALは表現するモデルであって，ここにINDSCALが単に軸の伸縮だけで個人差を説明するという，当該モデルのもつ限界が存在する。しかるに，特に，ストレス値が0.2より大きい年度データはINDSCALで無理に分析しないことが合理的対処であろう。これらの年度については，そのデータ構造が共通対象空間の軸の伸縮だけでは捉えられない独特の布置構造を持っていると見做すべきで，3年間のデータは別途，Kruskalの方法で解析すべきである。データマイニングの例外的データは排除すべきであるという原則に従うより(福田他，2001)，これらのデータを注意深く解析処理すべきであるという立場を筆者はとる。

そこで，1992年度，1995年度，1996年度のデータをそれぞれKruskalの方法で解析して得た布置をつぎに示す。これらのKruskalの方法による布置をINDSCALによる共通布置空間Xと比べると(図7，図8)，共通布置Xからの軸の伸縮だけでは得られない独特の布置構造を，これらの布置が実際に持っていることが視覚的に理解できる。

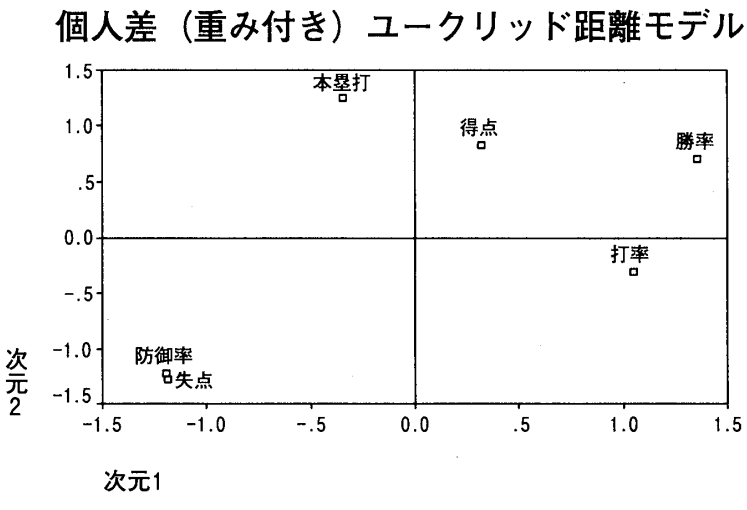

図６．INDSCALの９年間の共通布置（参考）

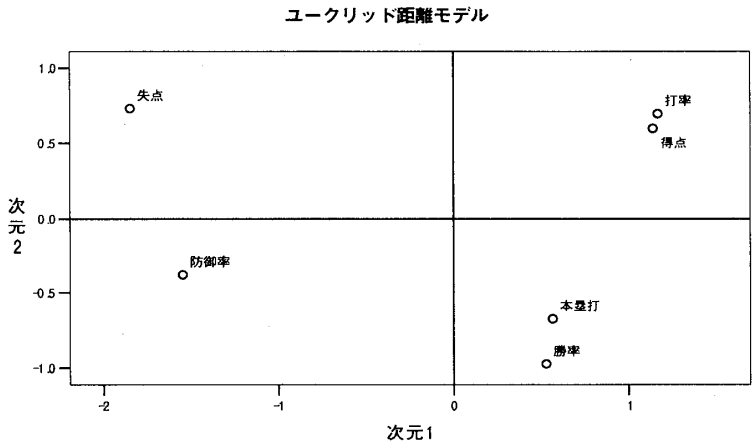

図７．1992年度の布置　ストレス値＝0.182

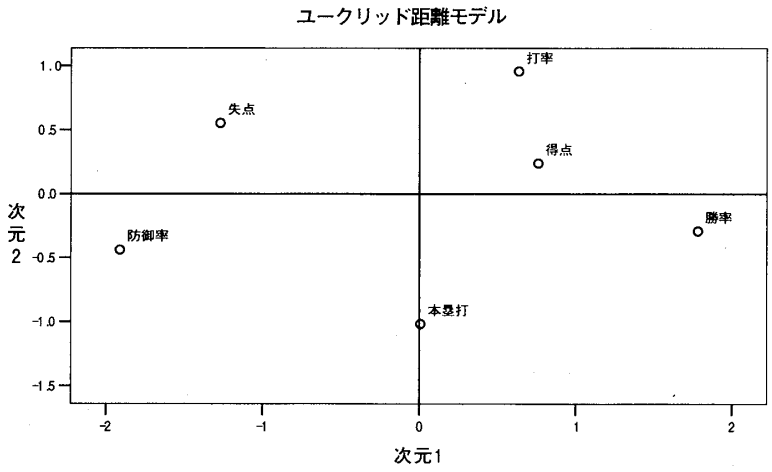

図８．1995年度布置　　ストレス値＝0.108

ところで，従来の統計的データ解析メソッドをデータマイニングの枠組みで使用するときには，使用形態に工夫を加える必要が一般的にある。例えば，データマイニングでクラスター分析を実行するときは，完全にｋ個のクラスターに分別する必要はなく，おおよその数のクラスターに分別すれば十分であり，例外的なデータを排除することで，その目的を達成することができる。福田らは(2001)データマイニングの実行時には例外的なデータや不要属性の排除を，データ解析メソッドを使用する場合の注意点として要求した。同様にデータマイニングで３元多次元尺度法を利用する際にも，例外的なデータは別途扱うことを本稿では実践した。結果として総括的なストレス値を0.166から0.119へ低下できたことで，効率的なデータ解析メソッドの有効利用を例証できた。

本稿は，従来はそれほど利用されなかった３元多次元尺度法INDSCALのデータマイニングへの適用が，データマイニング手法のクラスター分析や決定木と同様に有効なメソッドになり得ることを具体的なデータにINDSCALを適用して指摘した。特に当該モデルが３元データに対して安定した布置を供給するとともに，データ濃縮という事項の時系列変動を適切に捉えることを示し，また，データマイニングでのINDSCALの有効利用には例外的なデータは別途に取り扱うことが必要であることを強調した。

謝辞

本研究において流通経済大学のバンドン・エリ君，于永慶君には協力を頂きましたのでここに感謝します。

参考文献

Arabie, P., Carroll, J. D. & Desarbo, W. S. (1987). Three-way Scaling and Clustering. Newbery Park, CA: Sage.
岡太・今泉（共訳）(1990). 3元データの分析. 共立出版
Carroll, J.D. and Chang, J. J. (1970). Analysis of Indivisual Differences in Multi- Dimensional Scaling via an N-way Generalization of Eckart-Young Decompossion. *Psychometrika*, **35**, 283-319.
Everitt, B.S. & Rabe-Hesketh, S. (1997). The Analysis of Proximity Data. Kendall's Library of Statistics 4. Arnold, London.
Giudici, P. (2003). Applied Data Mining. Wiley, Chichester.
Kruskal, J. B. (1964). Nonmetric Multidimensional Scaling: Numerical Method. *Psychometrika*, **29**, 28-42.
Kruskal, J. B. (1965). Analysis of Factorial Experiments by Estimating Transformations of the Data. *Journal of the Royal Statistical Society Series B*, **27**, 251-263.
岡太彬訓，宮内綾子 (1996). INDSCALを用いた予測の一方法：新製品の購入予測心理学評論，30, 4，439-458.
奥喜正，前鶴政和 (2002). INDSCALのデータサイエンスへの適用と展望. 日本経営数学会第24回全国研究大会報告要旨集，27-31.
柴田　里程 (2001). データリテラシー.　共立出版
福田　剛志，森本　康彦，徳山　豪 (2001). データマイニング.　共立出版