原著論文

多次元データに対するユーザの主観的知識の可視化手法†

山本 康高*・吉川 大弘*・古橋 武*

ファジィ理論は科学に主観を取り入れる手法として注目されてきた。また、近年、チャンス発見をテーマとする研究において、ユーザの主観を利用することの重要性が指摘されている。本論文では、ユーザがデータに対して持つ主観的理解を可視化する手法を提案する。この可視空間は、自らの主観の内省及び、他者との知識の共有や議論を可能とする。提案手法では、ユーザがデータの特徴を説明していると思う言葉をリストアップし、その言葉に関連すると思う変数を選択する。それら変数を線形結合した射影軸を求め、可視空間を構築する。この線形射影軸の導出に主成分分析と正準判別分析を用いる。陸上の十種競技のデータに対する主観を可視化し、従来の解析法とは異なる新しい視点からデータを解釈できることを示す。さらに、十種競技における変数の全ての組み合わせに対して主成分分析法を適用した結果および、因子分析法を適用した結果との比較から、ユーザによる変数選択とその可視化の意義を明らかにする。

キーワード：主観、可視化、多次元データ、変数選択

1. はじめに

客観的であることを第一義として発展してきた科学に対して、ファジィは主観を取り入れることを提言してきた。菅野は、「ファジィ理論の行おうとしていることは科学の主観化である」と述べており[1]、その必要性を近代合理主義と比較して論じている。ファジィの独自性は、人が自らの主観に基づいて定めるメンバシップ関数によって言葉の意味の不確かさを表現できることにある。

一方、近年、チャンス発見[2]をテーマとして、人の主観を利用することの重要性が改めて注目されている。大澤は、チャンス発見を「意思決定において重要となる事象、状況またはそれらに関する情報を発見する活動」と定義し、「主観こそチャンス発見の本質的な鍵となる」と述べている[3]。また、KeyGraph[4]と呼ぶ、データの視覚化手法を提案し、得られた結果とユーザとの相互作用プロセスによるチャンス発見の可能性について議論している。

臼井ら[5]は、KeyGraphに主観的情報である触覚を取り入れる触視覚補完型KeyGraphを提案し、生地開発者による顧客の暗黙的ニーズの発見を行い、その有用性を示している。これは、主観を用いる情報技術

が実社会においても有効であることを証明した例といえる。この他にも、高間ら[6][7]は、免疫ネットワークモデルを用いたWeb情報の可視化手法を提案している。庄司ら[8]は、多次元尺度構成法[9]を用いたS-Conartと呼ぶシステムにより、ユーザのコンセプトの精緻化を支援する手法を提案している。

これらチャンス発見に関する手法の共通点は、情報を可視化する手法を用いていることである。可視化により、ユーザはデータに対する見方を内省でき、さらに、データに対する個人的視点を他者と議論することができる。すなわち、重要な情報を発見するための材料を可視化することにより、「知識の再確認」や「思いもよらない気づき」を期待するのである。

これらの手法では、主観は、結果を解釈しデータに対する「物語り」を想像するのに用いられている。大澤は、「目の前の出来事から物語りを作り出すことは重要であり[10]、上質な物語りは、グループの合意に達し、客観的にも正しいと評価されるようになる」と述べている[3]。このようにチャンス発見は、ファジィとは異なる主観の利用法を提供している。

本論文では、言葉によって表現されるユーザがデータに対して持つ主観的知識を可視化する手法を提案する。提案手法では、ユーザがデータの特徴を説明していると思う言葉をリストアップし、次にデータの持つ多くの変数の中から、その言葉に関連すると思う変数を選択する。このとき選択される変数は、ユーザの主観にまかせる。そして、それら変数の線形結合による射影軸を求め、これらの軸による可視空間を構築す

† A Visualization Method for User's Subjective View to Multi-Dimensional Data

Kosuke YAMAMOTO, Tomohiro YOSHIKAWA and Takeshi FURUHASHI

* 名古屋大学
Nagoya University

る。この場合、ユーザによりリストアップされた言葉は、データの主観的なラベル（以降、主観ラベルと呼ぶ）となる。ユーザによってデータに対する背景知識は異なり、同じ言葉に対しても選択する変数は異なる。提案手法はこの主観に基づいた選択の結果を可視化するものである。可視化されることで、ユーザは自らの主観を内省でき、さらに他者はその主観的知識を捉えて議論に供することができる。

本論文では、ユーザの主観ラベルに対して選択した変数を用いてデータを可視化するために、主成分分析（Principal Component Analysis：PCA）と正準判別分析（Multiple Discriminant Analysis：MDA [11]）を適用する手法について述べる。PCAを基礎とする手法は、データを要約した軸を同定する。またMDAは、与えられたデータのクラスを判別する軸を同定する。PCAやMDAは、線形射影軸により多次元のデータを情報圧縮できるため、可視化に有効である。また、線形射影は、その結合係数により、可視化結果と変数間の関係の把握を容易にし、ユーザの主観的知識の内省を図ることができると考えられる。

本論文のPCAを用いる手法において、射影軸を同定する手法は、グループ主軸法[12][13]の一種と捉えることができる。しかし、グループ主軸法においては変数の選択に基準はない[13]。本論文の手法は、ラベルの設定やラベルに対応する変数の選択をユーザの主観に基づいて行う点に特徴がある。一方、主観的なデータの捉え方の可視化により、主観の内省や他者との共有を可能にするという発想は、KeyGraphなどの手法と共通している。また、提案手法は菅野が主観に対して述べている「ファジィ理論においては、主観は解釈の手段ではなく、むしろ解釈すべき目的物である。」[1]という考えに通じるところがある。ただし、主観的知識の可視化を目指していることや、ユーザの主観に依存した変数選択により可視空間を構築することは、他の研究には見られない提案手法の独創点である。

本論文では、データとして陸上男子の十種競技の成績表（<http://ww1.usatoday.com/AT.asp?Event=ATM9001>）を用いて実験を行う。このデータに対して、主観ラベル（Subjective Label）によるPCA（以降、SL-PCA）を適用した結果を示す。また、十種競技における変数の全ての組み合わせに対してPCAを適用した結果と、因子分析の結果から、提案手法におけるユーザによる変数選択の意義を明らかにする。次に、主観ラベルによるMDA（以降、SL-MDA）と従来のMDAの結果とを比較し考察する。実験結果から、提案手法がユーザの主観を表現する有効な手法の

一つであることを示し、可視化したデータから、十種競技の特徴に対する知識の再確認や新しい仮説のための物語りを構築できることを示す。

以降、2章では提案手法について説明し、3章において実験結果および考察を示す。最後に4章でまとめる。

2. 提案手法

本手法の要点は、ユーザが主観的に言葉に関連すると思われる変数を選択し、その主観ラベルを線形射影して表現することにある。本論文では、データを要約する軸の同定にPCAを用いる手法と、また、データのクラスを判別する軸の同定にMDAを用いる手法を提案する。2.1節において一般的なPCAとMDAについて概説し、2.2節で提案手法について説明する。

2.1 主成分分析と正準判別分析

表1は、本論文で用いる記号と変数を示している。なおデータは、各入力変数において平均0、標準偏差1に標準化されているものとする。

a. PCA

主成分分析（PCA）は、複数の入力変数から互いに無相関な線形結合による座標変換を行い、情報を要約する手法である。PCAでは、射影軸の結合係数 w は(1)式に示す固有値問題の固有ベクトルとして得られる。このとき固有ベクトルは、固有値の大きな順に第1～ P 主成分と呼ばれる。また固有値は、各主成分上でのデータの分散に等しい。

表1 本論文で用いる記号一覧

i	クラス番号を示す添字
k	データ番号を示す添字
P	原空間の次元数
Q	圧縮後の次元数
N_D	データ数
N_C	クラスの数
n_i	i 番目のクラスに含まれるデータ数
$\mathbf{x}_k$	k 番目のデータ $(\mathbf{x}_{k1}, \mathbf{x}_{k2}, \dots, \mathbf{x}_{kP})^t$
W	射影軸からなる変換行列
w	射影軸の結合係数
S_B	クラス間散布行列
S_W	クラス内散布行列
m_i	i 番目のクラスの中心
$\mathbf{v}$	全データの平均
λ	ラグランジュ係数

$$S\mathbf{w} = \gamma\mathbf{w} \quad (1)$$

ここで、 γ は固有値を意味する。このとき、 S は入力変数の分散共分散行列であり、

$$S = \frac{1}{N_D - 1} \sum_k^{N_D} (\mathbf{x}_k - \mathbf{v})(\mathbf{x}_k - \mathbf{v})^t \quad (2)$$

によって得られる。 $\mathbf{v}$ は、全データの平均ベクトルを表し、

$$\mathbf{v} = \frac{1}{N_D} \sum_k^{N_D} \mathbf{x}_k \quad (3)$$

となる。

b. MDA

正準判別分析 (MDA) は、与えられた N_C 個のクラスを Q ($\leq \min(N_C - 1, P - 1)$) 本の判別式で分割することを目的とする。射影空間におけるクラス内散布行列 ($\mathbf{W}^t \mathbf{S}_W \mathbf{W}$) に対するクラス間散布行列 ($\mathbf{W}^t \mathbf{S}_B \mathbf{W}$) の比を最大にする変換行列 $\mathbf{W}$ を同定する。散布の程度を表すスカラー指標は散布行列の行列式となる。そのため、MDAにおける目的関数は、

$$J(\mathbf{W}) = \frac{|\mathbf{W}^t \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^t \mathbf{S}_W \mathbf{W}|} \quad (4)$$

と与えられる。このとき、クラス間散布行列 $\mathbf{S}_B$ およびクラス内散布行列 $\mathbf{S}_W$ は、

$$\mathbf{S}_B = \sum_i^{N_C} n_i (\mathbf{m}_i - \mathbf{v})(\mathbf{m}_i - \mathbf{v})^t \quad (5)$$

$$\mathbf{S}_W = \sum_i^{N_C} \sum_{\mathbf{x}_k \in N_{Di}} (\mathbf{x}_k - \mathbf{m}_i)(\mathbf{x}_k - \mathbf{m}_i)^t \quad (6)$$

と表される。ここで、(5) (6) 式中の各クラスの中心 ($\mathbf{m}_i$) は、

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x}_k \in N_{Di}} \mathbf{x}_k \quad (7)$$

となる。

(4) 式を最大にする判別式 ($\mathbf{W}$) を求めることは、

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w} \quad (8)$$

という一般化された固有値問題を解くことに等しい [11]。この解は特性多項式

$$|\mathbf{S}_B - \lambda \mathbf{S}_W| = 0 \quad (9)$$

により固有値 (λ) を求め、

$$(\mathbf{S}_B - \lambda \mathbf{S}_W) \mathbf{w} = 0 \quad (10)$$

を解くことで導かれる。

2.2 主観的知識の可視化手法

ユーザが持つデータを説明するために用いる主観ラベルと、それらに関連すると考える入力変数の対応表を表2に示すように作成する。表において、関連があると考えた箇所には○を付与する。このとき的主観ラベルと入力変数の対応をユーザが主観的に判断することが重要である。

この対応表に基づき、可視空間を形成するための線形射影軸を同定する。PCAおよびMDAに適用する手順を一括して以下に示す。また、主観ラベル (Subjective Label: SL) を用いた PCA/MDA であるため、それぞれ SL-PCA/SL-MDA と記す。

1. 主観ラベルと入力変数の対応表を作る。
2. 各主観ラベルに対して選択された入力変数のみを用いて PCA (MDA) を行う。
3. PCA (MDA) によって得られた最大の固有値を持つ固有ベクトルをその主観ラベルを表す射影軸として用いる。
4. 2あるいは3個の主観ラベルを選択して可視空間を構築する。

表2 主観ラベルと入力変数との対応表の例

入力変数	主観ラベル A	主観ラベル B	主観ラベル C
x_1	○	×	×
x_2	○	○	×
x_3	×	×	○
x_4	×	○	×
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_P	×	×	×

3. 実験と考察

本章では、提案手法の有効性を実験により示す。3.1節において、実験で用いたデータと実験内容を示す。3.2節では、SL-PCAを適用した結果を示し、変数の全ての組み合わせに対してPCAを適用した結果と因子分析の結果から、ユーザの主観による変数選択の意義を調べた実験結果について示す。また、3.3節でSL-MDAを適用した結果と従来のMDAについての比較結果を示す。

3.1 実験内容

多次元データとして、2000年シドニーオリンピックにおける陸上男子の十種競技の成績表を用いた、十種

表3 十種競技の内容

競技日程	競技名
一日目	100m 走
	走り幅跳び
	砲丸投げ
	走り高跳び
	400m 走
二日目	110mH(ハードル)
	円盤投げ
	棒高跳び
	槍投げ
	1500m走

表4 十種競技における筆者Aの主観ラベルと入力変数との対応表

競技名	スピード	力	持久力
100m 走	○	×	×
走り幅跳び	○	×	×
砲丸投げ	×	○	×
走り高跳び	×	×	×
400m 走	○	×	○
110mH	○	×	×
円盤投げ	×	○	×
棒高跳び	○	○	×
槍投げ	○	○	×
1500m走	×	×	○

競技は、表3に示す十種類の競技を二日間で行い、その記録を得点に換算して合計を競う陸上競技のことである。実験では、参加選手中、十種全ての競技に記録を残した24人分のデータを用いた。データの前処理として、各種目において値の大きい方が成績が良くなるように符合を変換し、平均0、標準偏差1となるように標準化した。

本実験では、筆者らの一人（以降、筆者A）が十種競技の成績は「スピード」、「力」、「持久力」という3つの要素が重要であると考えたため、これらを主観ラベルとしてデータを可視化した。筆者Aは、陸上競技に関する専門的知識はないものの、ほとんどの種目について実施経験があり、各種目に対して必要とされる能力に関する基本知識は備えていた。筆者Aは主観ラベルに関連する入力変数を表4のように選択した。

3.2 SL-PCAにおける結果と考察

SL-PCAにより得られた射影軸の結合係数と固有

表5 SL-PCAの固有値と結合係数

	スピード	力	持久力
固有値	2.68	1.71	1.15
100m 走	0.53	×	×
走り幅跳び	0.54	×	×
砲丸投げ	×	○	×
走り高跳び	×	×	×
400m 走	0.42	×	0.71
110mH	0.48	×	×
円盤投げ	×	○	×
棒高跳び	-0.04	○	×
槍投げ	0.15	○	×
1500m走	×	×	0.71

値を表5に示す。表中では、絶対値が0.3以上の結合係数は、各射影軸に対して影響が大きいと判断して太字にしてある。

表5より、SL-PCAにより、各主観ラベルに対してより関係が強いと考えられる競技の結合係数が大きく

表6 SL-PCAにより射影されたデータの座標

順位	スピード	力	持久力
1	2.90	0.67	2.05
2	2.06	1.13	0.98
3	3.20	1.82	0.16
4	2.55	0.30	2.73
5	1.37	0.09	-0.57
6	1.15	3.17	0.03
7	1.40	-1.52	0.65
8	-1.36	1.55	1.07
9	0.43	-0.57	0.09
10	0.51	0.22	-0.98
11	0.30	1.07	-0.57
12	-0.77	0.83	0.30
13	0.90	0.45	-1.16
14	-1.33	-0.43	1.18
15	-0.24	-0.48	-0.25
16	-1.37	0.66	-0.20
17	-1.34	-0.64	-0.92
18	-1.01	1.23	-1.46
19	-1.63	-1.38	-1.53
20	-1.52	-1.28	-0.75
21	-2.81	-1.25	1.10
22	-1.24	-1.39	-0.03
23	-2.24	-2.12	-1.06
24	0.07	-2.11	-0.86

なっていることがわかる。「棒高跳び」や「槍投げ」などは、このデータの範囲内では筆者Aの考える「スピード」とは相関が小さいことがわかる。また、「力」についても同様に「棒高跳び」の係数は小さくなっている。

表5の射影軸を用いて、3次元空間にプロットした結果を図1に示す。ここで、十種競技における上位と下位の成績者の差を明らかにするため、データを十種の総合成績の高い順に6人ずつA～Dの4つにランク分けしている。

表6は、表5に示す射影軸にデータを射影したときの座標を示している。見やすいように表中の負の値には網掛けがしてある。データを標準化してあるため、成績が平均以上である競技についてはデータの値が正となる。表6に着目すると、成績上位者が各軸において正の値をとり、下位の選手が負の値となる傾向を持つ。これにより筆者Aは、十種競技を要約して説明するための自らの知識が確からしいことが確認できる。また、それぞれのデータが可視空間上において説明可能であり、他者との知識共有が可能であるため、妥当な結論を導く議論が期待できる。これにより、SL-PCAによる可視空間は、筆者Aの主観的知識の表現系の一つになると考えられる。

通常、PCAでは、射影による情報の縮約度合いである寄与率を重視する。 j 番目の主成分に対する寄与率(C_j)は、対応する固有値を γ_j とすると次式により得られる。

$$C_j = \frac{\gamma_j}{\sum_{j'} \gamma_{j'}} \quad (11)$$

これにより、全データ中におけるどの程度の情報を、第 j 主成分が保持しているかを知ることができる。SL-PCAでは、選択された変数に対してPCAを適用した際の第1主成分の寄与率を各射影軸の縮約度とする。SL-PCAでは、射影軸が直交関係にないため、各軸の寄与率は全データに対する情報の縮約度の指標とはなり得ない。しかしながら、線形射影軸によりユーザの主観的知識を表現することを目指す本研究では、この値は選択された変数の情報をどの程度用いて被験者の主観的知識を表しているかの指標となる。SL-PCAにより得られた射影軸の縮約度を表7に示す。

表7から、各射影軸とも選択した変数の半分程度の情報が縮約されていることがわかる。この結果は、筆者Aが第1主成分に多くの情報を保持できる変数を主観的に選択していたと判断できる。

次に、ユーザによる変数選択の意義を実験的に検証

図1 SL-PCAによる可視化結果

表7 SL-PCAにおける縮約度

	スピード	力	持久力
縮約度	0.45	0.43	0.58

した結果について示す。主観ラベルに対して選択する変数はユーザに依存する。この選択結果が可視空間の構築に大きく影響するならば、提案手法は、ユーザの主観性に応じた可視空間を構築しているといえる。

そこで十種競技に対するすべての変数の組み合わせ(1023通り)に対してそれぞれPCAを適用し、可視空間を構築する線形射影軸と変数選択の関係について調査した。各変数の組み合わせに対してPCAを適用後、射影軸に与える影響が小さいと判定された変数はその組み合わせから削除した。この変数の削除は次のようにして行った。PCAを適用した際の第一主成分の係数において、絶対値が0.3以下であった変数を削除し、残った変数に対して再度PCAを適用し、削除される変数がなくなるまで、同様の処理を繰り返した。この方法を、例えば、筆者Aがスピードという主観に対して選んだ変数(表4、5)に適用すると、100m走、走り幅跳び、400m走、110m Hの4つの変数が残る。

この実験において、初期の変数の組み合わせに依存せず、特定の変数の組み合わせに収束するのであれば、その組み合わせはデータを説明するために重要な情報であると考えられる。しかし、実験の結果1023通りの変数の組み合わせから、208通りもの組み合わせに収束する結果を得た。短距離と砲丸投げの係数が残りやすい傾向はあったものの、ある特定の変数の組み合わせへの収束は見られなかった。これは人間による変数選択が、射影軸において重要な役割を果たしたことを意味している。すなわち、ユーザの変数選択が可視化されたデータ分布に反映され、ユーザの主観的知識が可視化されたとする重要な根拠の一つが得られ

た。

次に、データの統計的性質を基に解析を行う手法により、主観ラベルと同様の射影軸が得られるか実験を行った。実験には、射影軸を回転することにより解釈を容易にできる因子分析法を用いた。因子抽出に主因子法を用い、バリマックス回転を行ったときの因子負荷量を表8に示す。データを説明する潜在変数とみなせる固有値が1以上の因子を4つ同定できた。

第1因子は短距離や跳躍に関する競技の係数が高く、選手のバネを表していると解釈した。第2因子は、全てではないものの、力と走力を必要とする競技の係数が高いため、選手の総合力と解釈した。第3因子は1500m走の係数が高いため、持久力と解釈でき、第4因子は言葉によって説明するのが困難な軸となった。重要なことは、因子分析法により、解釈容易な射影軸を得ているが、その解釈が主観ラベルとは異なることである。因子分析では、因子が不定性を持ち、解釈を容易にするための回転が行えるため、発見的に分析を繰り返すことにより、様々な結論を導くことができる。また解析結果に対する解釈は解析者の知識や経験に依存するため、上記以外の解釈も考えられる。しかしながら、因子分析は基本的に相関行列に対する解析であり、相関の高い変数が共通の因子に対して大きな因子負荷量を持つ傾向がある。回転は、共通の変数が複数の因子に大きな因子負荷量を持たないようにするなどの単純構造原理に基づいて行われ、因子負荷量における各行あるいは各列の分散値などを基準とする手法が一般的である。そのため、データの統計的性質に基づく解析結果が得られることになる。故に上述のPCAの結果に見られたように、統計的性質だけでは得ることが困難な主観ラベルを基としたデータの解釈が行えるとは限らない。

因子分析には、統計的性質を基にデータを解釈でき

表8 因子分析における因子負荷量

	第1因子	第2因子	第3因子	第4因子
100m走	0.59	0.63	-0.22	0.25
走り幅跳び	1.00	0.09	0.07	-0.14
砲丸投げ	0.25	0.88	-0.14	-0.08
走り高跳び	0.46	0.00	0.15	-0.11
400m走	0.47	0.41	0.26	0.65
110mH	0.68	0.22	-0.41	-0.03
円盤投げ	-0.07	0.69	0.10	-0.11
棒高跳び	-0.09	-0.07	-0.06	0.37
槍投げ	0.25	0.12	0.28	-0.41
1500m走	0.03	-0.04	0.95	-0.13

るという有用性があり、提案手法には、ユーザが主観ラベルとそれに対する変数選択を行うことにより、自らの知識の内省および他者との共有ができるという有用性があると考えられる。これらの手法は、排他的ではないため、必要に応じて使い分けることが望ましい。

3.3 SL-MDAにおける結果と考察

3.2節で示したA～Dの4つの各ランクをそれぞれ1つのクラスとみなし、SL-MDAを適用した。これにより、筆者Aがクラスを分けるために用いている主観を可視化することになる。また、比較として従来のMDAを適用した結果についても示す。従来のMDAでは、各ランクを最も分ける判別式を得ることができる。

SL-MDAにより得られた射影軸の結合係数と固有値を表9に、従来のMDAにおいて得られた固有値と判別式を表10に示す。表中では絶対値が0.3以上の結

表9 SL-MDAの固有値と結合係数

	スピード	力	持久力
固有値	7.47	3.03	0.76
100m走	0.34	×	×
走り幅跳び	-0.15	×	×
砲丸投げ	×	0.74	×
走り高跳び	×	×	×
400m走	0.53	×	1.00
110mH	0.36	×	×
円盤投げ	×	0.17	×
棒高跳び	-0.16	0.14	×
槍投げ	0.65	0.64	×
1500m走	×	×	0.08

表10 従来のMDAの固有値と判別式

	第1判別式	第2判別式	第3判別式
固有値	42.19	0.80	0.23
100m走	0.01	-0.06	-0.16
走り幅跳び	-0.07	-0.76	-0.45
砲丸投げ	0.33	0.28	-0.03
走り高跳び	0.33	0.11	0.40
400m走	0.53	0.10	0.63
110mH	0.35	0.32	0.15
円盤投げ	0.18	-0.26	-0.43
棒高跳び	-0.05	0.12	-0.06
槍投げ	0.58	0.36	-0.01
1500m走	0.02	0.05	-0.10

合係数を太字にしてある。

表9では、筆者Aが選手のランクを判別すると考えた射影軸が同定されている。SL-PCAのときと同じ主観ラベルを用いているものの、得られる射影軸の係数は異なる。その結果、例えば筆者Aがスピードという言葉に対して持っている主観的理解は、データを要約するときと判別する時とは異なって可視化されている。

一方、表10の固有値より、MDAでは第1判別式が4つのランクを分けるのに大きく寄与していることがわかる。また、結合係数の大きな競技が多いことから、この判別式は「総合力」と解釈してよさそうである。

表11, 12は、それぞれ表9, 10に示す射影軸にデータを射影したときの座標を示している。表中の負の値には網掛けがしてある。これらの値を3次元空間にプロットした結果を、図2, 3に示す。

図2から、提案手法では、各軸がそれぞれの主観ラベルの観点から4つのランクを分割しており、結果として3次元の可視空間の対角線上にデータが分布して

いることがわかる。すなわち、この可視空間においては、上述の対角線が選手の総合力を表すことになる。この他にも、「スピード」と「力」がランクを分ける重要な要素であり、どれか1つの能力だけが高いのでは、ランクAの成績を収めることが困難であることがわかる。これは、筆者Aの持つ十種競技に対する知識を可視空間により再確認できる結果であるとともに、明示性が高く説明が容易であることから、他者との議論においても有用であるといえる。また、破線で囲んだ選手aは「持久力」をつけること、選手bは「力」をつけることが、金メダルへの課題であるという物語りを、筆者Aの主観を通して構築できる。

従来のMDAは図3より、4つのランクが第1判別式のみによって分割されていることが視覚的にも把握できる。MDAは、クラス間/内のデータ散布を基準として、最も良い判別式を同定するため、提案手法よりも、異なるクラスのデータの混在が少なく、しかしながら、提案手法では、筆者Aの主観を可視化することによって、知識の再確認や新たな発見等の従来の

図2 SL-MDAによる可視化結果

図3 従来のMDAによる可視化結果

表11 SL-MDAにより射影されたデータの座標

順位	スピード	力	持久力
1	2.90	0.67	2.05
2	2.06	1.13	0.98
3	3.20	1.82	0.16
4	2.55	0.30	2.73
5	1.37	0.09	-0.57
6	1.15	3.17	0.03
7	1.40	-1.52	0.65
8	-1.36	1.55	1.07
9	0.43	-0.57	0.09
10	0.51	0.22	-0.98
11	0.30	1.07	-0.57
12	-0.77	0.83	0.30
13	0.90	0.45	-1.16
14	-1.33	-0.43	1.18
15	-0.24	-0.48	-0.25
16	-1.37	0.66	-0.20
17	-1.34	-0.64	-0.92
18	-1.01	1.23	-1.46
19	-1.63	-1.38	-1.53
20	-1.52	-1.28	-0.75
21	-2.81	-1.25	1.10
22	-1.24	-1.39	-0.03
23	-2.24	-2.12	-1.06
24	0.07	-2.11	-0.86

表12 従来のMDAにより射影されたデータの座標

順位	第1判別式	第2判別式	第3判別式
1	1.86	-0.52	0.13
2	2.02	0.11	0.31
3	1.40	-1.20	-0.86
4	1.70	-0.94	0.89
5	1.44	0.42	1.05
6	1.60	-0.17	-1.19
7	0.46	0.42	0.66
8	0.75	0.34	-0.12
9	0.56	0.75	0.40
10	0.57	0.96	-0.08
11	0.49	0.11	-0.18
12	0.31	0.77	0.25
13	-0.24	0.12	-0.42
14	-0.75	-0.09	-0.17
15	-0.29	-0.33	-0.13
16	-0.46	0.40	-0.22
17	-0.76	-0.09	-0.57
18	-0.46	0.21	-1.16
19	-1.55	-0.18	-0.52
20	-1.43	-0.27	-0.06
21	-1.85	0.61	-0.21
22	-1.57	-0.71	0.92
23	-2.01	-0.31	0.43
24	-1.80	-0.43	0.84

MDAとは異なる情報を得ることができた。

また、表11, 12に着目すると、SL-PCAと同様、両手法とも成績上位者が各軸において正の値をとり、下位の選手が負の値を持つ傾向を持つことがわかる。MDAが第1判別式で総合能力を分類しているのに対して、SL-MDAが、筆者Aの主観的観点から、選手の能力を3つに分解し可視化できたと考えられる。

次にSL-MDAにおけるクラスの分離の精度の評価方法について説明する。通常、MDAでは、 j 番目の判別軸におけるクラスの分離度合いを意味する相関比(η_j^2)が、次式により表される。

$$\eta_j^2 = \frac{\lambda_j}{1 + \lambda_j} \quad (12)$$

ここで、 λ_j は j 番目の判別軸の固有値を意味している。この指標は0~1の値をとり、1に近いほど、クラスの分離性が高いことを意味する。SL-MDAにおいても、この指標を用いることにより、線形射影軸として表現したユーザの主観的知識によるクラスの分類度合いを定量化する。表13に、MDAの各判別式にお

表13 MDAのSL-MDAの相関比

MDA	第1判別式	第2判別式	第3判別式
相関比	0.98	0.44	0.19
SL-MDA	スピード	力	持久力
相関比	0.88	0.75	0.43

ける相関比と、SL-MDAの各主観ラベルに対する相関比を示す。

表13のMDAの相関比に着目すると、第1判別式が高い精度でクラスを分離できていることが定量的に確認できる。また、SL-MDAの相関比からは、特にスピード、力においてクラスの分離性が高いことを確認できる。これによりユーザは自らの主観的知識の観点から十種競技の選手の能力をどの程度判別できるかを定量的に知ることが可能となる。

以上の実験結果より、主観ラベルを用いる提案手法によって、ユーザの主観を可視化でき、主観的知識の再確認や物語りを生み出すのに利用できる可能性が示された。ユーザによる変数選択の有効性は、常に正しい結果を導くとは限らない。しかしながら、射影軸同定の変数選択への依存性の強さや、主観性や解釈性の定量化が困難であることから、ユーザ自らが変数を選択し、それを基に可視化を行うことは、有効なデータ解釈法の一つになると期待される。

4. おわりに

本論文では、人の主観を利用することの重要性が再認識されていることに着目し、ユーザの持つデータに対する主観的知識を可視化する手法を提案した。本手法では、ユーザがデータの特徴を示していると思う言葉(主観ラベル)に関連すると判断した変数に対してPCA/MDAを適用することで可視空間を構築する。本手法を十種競技のデータに対して適用し、筆者らの一人の主観的知識を可視化した。PCAを基礎とした提案手法により得られた結果と、すべての変数の組み合わせに対してPCAを適用した結果との比較から、射影軸の変数選択への依存性を示し、ユーザによる変数選択が有効なデータ解釈の手段の一つになりうることを確認した。同データに対して因子分析法を適用し、同定された因子の解釈は提案手法とは異なることを示した。また、MDAを基礎とした提案手法により得られた結果より、従来のMDAとは異なる視点からデータの解釈ができ、ユーザの知識の再確認、また十種競技者の技術向上のための物語りの構築ができることを示した。

本論文では、十種競技の記録という客観的なデータ

を取り扱っているものの、人の感性を定量化する Semantic Differential (SD) 法や尺度評定法などにより得られる主観データに対しても、提案手法が適用可能である。主観データの解析において、各個人が言葉に対して抱く印象の差を反映させる必要がある。このようなデータに対して提案手法では、印象の差を主観的変数選択により表現できると考えられる。今後、主観データに対する提案手法の有用性についても検証していきたい。また、射影軸上におけるファジィ集合を考慮した主観の同定法について検討していきたい。

謝辞

本研究の一部は日本学術振興会特別研究員奨励費、および文部科学省科学研究費基盤研究C(2)No.16500126の補助金を得て遂行された。

参考文献

- [1] 菅野道夫, ファジィ理論の展開, サイエンス社, 1989.
- [2] 大澤幸生, チャンス発見の情報技術, 東京電機大学出版局, 2003.
- [3] 大澤幸生, “チャンス発見の3つの原理,” 日本知能情報ファジィ学会誌, vol.15, no. 3, pp.260-269, 2003.
- [4] 大澤幸生, Nels E. Benson, 谷内田正彦, “Key-graph: 単語共起グラフの分割・統合によるキーワード抽出,” 電子情報通信学会論文誌, vol.J82-D1, no.2, pp.391-400, 1999.
- [5] 臼井優樹, 大澤幸生, “組織内合意形成のためのチャンス発見-生地メーカーにおける触視覚補完型 KeyGraph の効果-,” 日本知能情報ファジィ学会誌, vol.15, no.3, pp.275-285, 2003.
- [6] 高間康史, 廣田薫, “免疫ネットワークメタファに基づく Web 情報可視化手法,” 日本ファジィ学会誌, vol.14, no.5, pp.472-481, 2002.
- [7] 高間康史, “Softにおけるチャンス発見-投稿論文データの情報可視化から-,” 日本知能情報ファジィ学会誌, vol.15, no.3, pp.270-274, 2003.
- [8] 庄司裕子, 堀洪一, “購買におけるコンセプト精緻化を支援するためのインタラクション手法とその評価,” 日本知能情報ファジィ学会誌, vol.15, no.3, pp.297-308, 2003.
- [9] J.W. Sammon, “A nonlinear mapping for data structure analysis,” IEEE Trans. on Computers, vol.C-18, no.5, pp.401-409, 1969.
- [10] H. Predinger, and M. Ishizuka, The storification of chances. Corporate training with life-like characters in virtual social environments, in Ohsawa, Y. and McBurney, P. (eds) Chance Discovery, Springer, 2003.
- [11] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification-2nd ed., A Wiley., 2000.
- [12] 渡部 洋編著, 心理・教育のための多変量解析入門, 福村出版, 1992.
- [13] 芝祐順, 行動科学における相関分析法, 東京大学出版会, 1975.

(2004年11月15日 受付)
(2005年12月04日 採録)

問い合わせ先

〒464-8603 名古屋市 千種区 不老町
名古屋大学大学院 工学研究科 計算理工学専攻
山本 康高
TEL: 052-789-2793
FAX: 052-789-3166
E-mail: yamamoto@cmlpx.cse.nagoya-u.ac.jp

著者紹介

やまもと こうすけ
山本 康高

1978年11月9日生まれ。2003年3月三重大学大学院 工学研究科 電気電子工学専攻 博士前期課程修了。同年4月、博士後期課程に進学。2004年4月より名古屋大学大学院 工学研究科 計算理工学専攻に転入学。現在に至る。多次元データの可視化技術および感性データ解析に関する研究に従事。

よしかわ ともひろ
吉川 大弘

1970年7月17日生まれ。1997年名古屋大学大学院博士後期課程修了。同年カリフォルニア大学バークレー校ソフトコンピューティング研究所客員研究員。1998年三重大学工学部電気電子工学科助手。2005年名古屋大学工学研究科 COE 特任助教授。現在に至る。主としてソフトコンピューティングとその応用に関する研究に従事。博士(工学)。

ふるはし たけし
古橋 武

1985年3月名古屋大学大学院博士課程修了。1990年12月より同大工学部助教授。2001年1月より三重大学 情報工学科 教授。2004年4月より名古屋大学大学院工学研究科教授。現在に至る。工学博士。94年日本ファジィ学会論文賞受賞。主として、ソフトコンピューティングに関する研究に従事。

A Visualization Method for User's Subjective View to Multi-Dimensional Data

by

Kosuke YAMAMOTO, Tomohiro YOSHIKAWA and Takeshi FURUHASHI

Abstract :

Fuzzy theory has attracted a lot of attention as a method that introduces human subjectivity into science. Recently, the importance of utilization of a user's subjective view has been pointed out in the field of Chance Discovery. This paper proposes a method that visualizes the user's subjective view of multi-dimensional data. The user lists words that he/she considers appropriate to explain some of the features of data, and selects variables associated with each word based on his/her view. A visualized space is constructed with projection axes that are obtained as linear combinations of the selected variables for each word. The proposed method employs the Principal Component Analysis (PCA) or Multiple Discriminant Analysis (MDA) for identifying the projection axes. Through the visualized data, the user can observe his/her own subjective view of the data, and he/she can share and discuss it with other people. This paper applies the proposed method to a data set of decathlon in the Sydney Olympic Games. In this experiment, the user's subjective view of data is visualized. This paper shows that the proposed method succeeds in providing the user with a new interpretation of the data different from the conventional ones. This paper also shows that the visualization of the data using subjective variable selection can yield meaningful results different from those of the PCA applied to all the combinations from ten variables, and that of factor analysis.

Keywords : Subjectivity, Visualization, Multi-dimensional data, Variable Selection

Contact Address : **Kosuke YAMAMOTO**

Nagoya University

TEL : 052-789-2793

FAX : 052-789-3166

E-mail : yamamoto@cmplx.cse.nagoya-u.ac.jp