ファジィクラスタリングを用いた主成分分析における 変量選択とデータマイニングへの応用*

本多 克宏[†]・市橋 秀友[†]

Variable Selection in PCA Using Fuzzy Clustering and Its Application to Data Mining*

Katsuhiko HONDA[†] and Hidetomo ICHIHASHI[†]

Knowledge discovery in databases (KDD) or data mining involves fitting models to or determining patterns from high dimensional data sets, and extraction of correlation rules plays an important role. This paper proposes a new approach to knowledge discovery with linear model estimation, in which principal component analysis (PCA) is performed by selecting variables. The proposed algorithm is a hybrid of fuzzy clustering and PCA based on lower rank approximation of data matrix, in which the relative responsibilities of the variables are estimated by using possibilistic constraint for memberships. The proposed algorithm is also enhanced to a local PCA model that can be used for data mining by performing both of linear model estimation and stratified sampling.

1. はじめに

大規模データベースからの知識発見 (Knowledge Discovery in Databases: KDD) やデータマイニングの手法として、データ集合に内在するルールの抽出がしばしば行われる。上田はそのための重要な要素として、「四つのS: 測定, 視覚化, 相関, 層別」を挙げている [1]。すなわち、測定されたデータ集合に層別を通して適切な前処理を施したうえで、潜在的な相関関係を見つけることが目的であり、人間にとってわかりやすく示すために視覚化の手法も用いられる。本論文では、データ集合に内在する相関関係の抽出法として、主成分分析を用いた線形モデルのあてはめを取り上げ、変量の選択を通じて相関関係を表現するための手法を提案する。

主成分分析における変量選択の手法はこれまで多くの提案があり、それらが森らの研究グループによりまとめられている [2]。これらの手法では、主成分分析の目的をデータ行列の情報圧縮における復元誤差最小化と位置づけ、部分空間上でのデータ配置を構成するための情報として冗長な変量を削除することで変量の選択を行っている。Jolliffe [3] は小さな固有値に対応する主成分への寄

与の大きい変量を不要な変量とみなして順に削除する方法を、Tanaka and Mori [4] は拡張主成分分析の規準を利用した変数選択法を、それぞれ提案している。しかし、これらの手法では相関の大きい変量ほど冗長な変量とみなされ、いずれかが削除されることとなるため、データマイニングを目的とした利用には不適である。そこで、以下ではファジィクラスタリング手法を利用して相関のある変量を選択することを考える。教師なし分類の手法であるクラスタリングをデータ解析手法と融合することで知識発見に利用する試みがいくつか提案されている。クラスタに含まれるデータの中心 (平均) をクラスタの代表 (プロトタイプ) とする k-means クラスタリング法にファジィメンバシップを導入した Fuzzy c-Means (FCM) 法 [5] は、プロトタイプを線形多様体に拡張することで Fuzzy c-Varieties (FCV) 法 [6] として線形構造の抽出に応用されているが、そのアルゴリズムは主成分分析に混合分布モデルの考えを導入した局所的な主成分分析 [7,8] と類似しており、データマイニング手法として用いられる [9,10]。また、FCV 法は主成分分析における最小 2 乗基準 [11] に、個体を分類するメンバシップを導入したモデルと同等とみなすことができ、局所的な主成分分析における欠測値や要素ごとの外れ値を処理する手法が開発されている [12,13]。一方、質的なデータの主成分分析ともみなされる等質性分析 [14] に項目を分類するメンバシップを導入することで、関連のある項目の群

* 原稿受付 2005 年 1 月 31 日

[†] 大阪府立大学 大学院 工学研究科 Graduate School of Engineering, Osaka Prefecture University; 1-1 Gakuen-cho, Sakai city, Osaka 599-8531, JAPAN

Key Words: principal component analysis, fuzzy clustering, data mining.

を見つける方法が提案されている [15]。これらの手法では対象をいくつかのクラスに分類する際に、全クラスへのメンバシップの総和が1となる確率的な制約が課されることから、混合分布モデルを用いた確率的な分類と類似した解釈が可能となる。すなわち、各対象の分類は競合的であり、メンバシップは相対的な帰属度を表すものであるが、クラスタリングにおけるメンバシップの制約条件としては、各対象の絶対的な帰属度を表そうとする可能性的な制約 [16] も提案されている。

そこで本論文では、主成分分析において変量を分類するためのメンバシップを導入することで、変量選択を行うファジィクラスタリングモデルを提案する。文献 [12,13] では、データ行列の要素ごとの近似に基づく主成分分析の目的関数に個体のメンバシップを導入していたが、これを変量のメンバシップに変更することで、線形モデルの推定における各変量の重要度を調整し、近似誤差の大きい変量群を無視し、誤差の小さい変量群をより重視してモデルの推定を行う。また、変量の絶対的な重要度を求めるために、可能性的な制約を用いる。ただし、可能性的な制約を用いる場合には初期分割に依存して結果が異なることが多いため、初期分割の影響を受けにくい段階的可能性クラスタリング法 [17] を採用する。

さらに、最小2乗基準を用いた線形ファジィクラスタリングと同様の枠組みで定式化されている提案法は、個体と変量の分割を同時に考慮したモデルに拡張することができる。すなわち、主成分分析における最小2乗基準に2種類のメンバシップを同時に組み込むことができ、局所的な主成分分析における変量の分類を可能とする。これはデータマイニングにおける主要な要素であるデータの層別と相関の抽出を同時に考慮することを意味する。

数値実験では、人工的に作成されたデータを用いて提案手法による変量選択を従来法と比較した後に、実世界のデータを用いてデータマイニングツールとしての有効性を検証する。

2. 主成分分析による情報縮約とファジィクラスタリング

2.1 主成分分析における変量選択

n 個体についての m 次元観測ベクトル $\tilde{\mathbf{x}}_i$, $i=1, \dots, n$ が与えられた際に、主成分分析による情報縮約を考える。観測値はデータ行列 $X=(x_{ij})$ とまとめて表現することもある。以下、太字は縦ベクトルであり、行列の行方向の要素からなるベクトルの場合は $\sim$ を付して表すものとし、 $X=(\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n)^\top$, または、 $X=(\mathbf{x}_1, \dots, \mathbf{x}_m)$ のように記述する。主成分分析は多変量からなるデータをなるべく情報の損失を少なくしながらデータの縮約を行う手法と定義される。ここでは、 m 次元ベクトル $\tilde{\mathbf{x}}_i$ を p 次元の主成分 $\tilde{\mathbf{f}}_i = A^\top(\tilde{\mathbf{x}}_i - \mathbf{b})$ に縮約するものとする。ただし、 $A=(\mathbf{a}_1, \dots, \mathbf{a}_p)$ は主成分ベクトル $\mathbf{a}_k$, $k=1, \dots, p$ を並べた主成分行列であり、平均値ベクトルを $\mathbf{b}$ とお

た。情報の損失量を最小とするデータの縮約を行うため、以下の最小2乗基準 [11] の最小化が図られる。

$$L_{lsc} = \text{tr} \left\{ (X - Y)^\top (X - Y) \right\} \quad (1)$$

ここで、 $Y = FA^\top + \mathbf{1}_n \mathbf{b}^\top$ であり、 $F=(\tilde{\mathbf{f}}_1, \dots, \tilde{\mathbf{f}}_n)^\top$ はおのおのの標本データの主成分得点からなる $(n \times p)$ の成分得点行列、 $\mathbf{1}_n$ はすべての要素が1である n 次元ベクトルである。(1) 式の最小化はデータ行列 X を最小2乗の意味で低階数近似する行列 Y を求める問題となる。ただし、唯一の解を求めるために、 $F^\top F = I$, $F^\top \mathbf{1}_n = 0$ および $A^\top A$ が対角行列となる制約条件を課される。

主成分分析における変量の選択法として、データの特徴を測りとれる指標を少ない次元で作成することを目指した手法が提案されている。すなわち、ある変量を削除する前後で主成分得点の配置に変化がほとんど無いような場合に、その変量を冗長な変量とみなして削除しようとするものである。たとえば、Fig. 1 のような3次元データが与えられた場合を考えてみよう。三つの変量のうち、 x_1 と x_2 は互いに相関が高いのに対して、 x_3 はそれらとは関連がないデータとなっている。このデータに主成分分析を施すと、破線で示された2次元空間上への射影として主成分が与えられる。この2次元空間上での配置を再現するうえでは、 x_1 と x_2 の持つ情報は冗長であり、いずれか一方があれば十分であるとみなして、 x_1 または x_2 を削除しようとするのが、従来の変量選択法である。

Fig. 1 Example of variable selection in PCA

しかし、データマイニングにおいて相関の高い変量の組を選択する問題を考えた際には、むしろ x_3 が無意味な変量であり、 x_1 と x_2 に特化することで両者の間の1次元構造を抽出することができるとも考えられる。以下では、ファジィクラスタリング手法を導入することで、変量の重要度を表す重みを考慮した線形モデルの推定法を提案する。

2.2 主成分分析とファジィクラスタリング

代表的なファジィクラスタリング法である Fuzzy c-Means (FCM) 法では、データを C 個のクラスに分類する際に、所属する標本データの中心 (平均) $\mathbf{b}_c$ をク

ラスタのプロトタイプとし、クラスタ中心からの 2 乗距離 d_{ci}^2 の総和が最小となるように分類を行う。そのため、以下の目的関数の最小化が図られる。

$$L_{fcm1} = \sum_{c=1}^C \sum_{i=1}^n u_{ci}^\theta d_{ci}^2 = \sum_{c=1}^C \sum_{i=1}^n u_{ci}^\theta (\tilde{\mathbf{x}}_i - \mathbf{b}_c)^\top (\tilde{\mathbf{x}}_i - \mathbf{b}_c) \quad (2)$$

ここで、 u_{ci} は第 i 標本データが第 c クラスタに帰属する度合いを表すメンバシップであり、 $\sum_{c=1}^C u_{ci} = 1$ なる確率的な制約が課される。k-means クラスタリングでは第 i 標本データが帰属するクラスタが唯一に与えられるため、 $u_{ci} \in \{0, 1\}$ と定められるが、ファジィクラスタリングでは帰属にあいまいさが許容されるために $u_{ci} \in [0, 1]$ で与えられる。(2) 式は $\theta = 1$ のときには u_{ci} について線形計画問題となり解が端点 ($u_{ci} \in \{0, 1\}$) で得られるため、目的関数を非線形関数に変換する標準的な FCM 法では $\theta > 1$ と定められる。また、 θ が大きいほど標本の所属があいまいとなり、 $\theta \rightarrow \infty$ のとき $u_{ci} \rightarrow 1/C$ となる。このように、 θ が分割のファジィ度を制御しており、fuzzifier とよばれる。クラスタリングアルゴリズムはクラスタ中心とメンバシップを交互に最適化する繰り返しアルゴリズムが用いられる。

また、(2) 式で $\theta = 1$ とおいたものに非線形項を付加することでファジィ分割を得ようとする正則化 FCM 法も提案されている。宮本ら [18] は (負の) エントロピー項を付加して、

$$L_{fcm2} = \sum_{c=1}^C \sum_{i=1}^n u_{ci} d_{ci}^2 + \lambda \sum_{c=1}^C \sum_{i=1}^n u_{ci} \log u_{ci} \quad (3)$$

なる目的関数を用いるエントロピー正則化 FCM 法を提案している。 λ が標準的な手法における θ と同様にファジィ度を制御する係数であり、 λ が大きいほどあいまいな分割となる。エントロピー正則化法を用いた場合も標準的な手法と類似したファジィ分割が得られるが、分類関数の形状に違いがある [18]。

FCM 法の拡張として、プロトタイプを中心点から p 次元の線形多様体に変更することで直線や平面状のクラスタを得ようとする Fuzzy c-Varieties (FCV) 法 [6] が提案されている。FCV 法では、(2) 式の d_{ci}^2 をデータ点と線形多様体との距離

$$d_{ci}^2 = \|\tilde{\mathbf{x}}_i - \mathbf{b}_c\|^2 - \sum_{k=1}^p |\mathbf{a}_{ck}^\top (\tilde{\mathbf{x}}_i - \mathbf{b}_c)|^2 \quad (4)$$

に変更される。ただし、 $\mathbf{a}_{ck}$, $k=1, \dots, p$ がプロトタイプを張る単位ベクトルで、互いに直交するものとする。このベクトル $\mathbf{a}_{ck}$ はファジィ散布行列の固有ベクトルとして算出されることから、FCV 法はしばしばクラスタごとにファジィ主成分分析 [19] を施す手法であるとみなさ

れる。また、FCV 法はエントロピー正則化を用いた場合には、Hinton ら [7] の局所的な主成分分析に等しく、データマイニング手法としても有効である [9, 10]。

本多ら [12, 13] は局所的な主成分分析への異なる定式化として、最小 2 乗基準にファジィクラスタリングのメンバシップを導入する手法を提案し、FCV 法と等しい解が得られることを示している。FCV 法がデータ点と p 次元線形多様体との距離をクラスタリング基準としているのに対して、最小 2 乗基準では行列の要素ごとの近似として定式化されるため、欠測値やサンプル内ノイズの影響を無視した局所的な主成分分析が可能であり、以下の目的関数が提案されている。

$$L_{rfcv} = \sum_{c=1}^C \sum_{i=1}^n u_{ci} \sum_{j=1}^m w_{cij} \left(x_{ij} - \sum_{k=1}^p f_{cik} a_{ckj} - b_{cj} \right)^2 + \lambda \sum_{c=1}^C \sum_{i=1}^n u_{ci} \log u_{ci} \quad (5)$$

ここで、ファジィ分割を得るためにエントロピー正則化が用いられている。また、 w_{cij} が第 c クラスタにおける要素 x_{ij} の重みであり、すべての重みを 1 とした場合には FCV 法と等しくなる。 x_{ij} が欠測値の場合は $w_{cij} = 0$ とおくことで、欠測値のみを無視して局所的な主成分分析が行われる [12]。M-推定に基づいてノイズとみなされる要素の重みを小さく与えた場合には、サンプル内ノイズにロバストな局所的主成分分析となる [13]。ただし、いずれの手法においてもファジィ分割のための重みは個体のメンバシップ u_{ci} のみであり、 w_{cij} はクラスタリング基準以外により定められている。

3. ファジィクラスタリングを用いた主成分分析における変量選択

3.1 変量選択のためのファジィクラスタリング

本節では、(1) 式の最小 2 乗基準に変量選択のためのメンバシップを導入したモデルとして、以下の目的関数の最小化問題を提案する。

$$L_{pca_v} = \sum_{i=1}^n \sum_{j=1}^m v_j \left(x_{ij} - \sum_{k=1}^p f_{ik} a_{jk} - b_j \right)^2 + \lambda_v \sum_{j=1}^m v_j \log v_j \quad (6)$$

ここで、 v_j が第 j 変量の重要度を表すメンバシップとし、ファジィ分割を得るためにエントロピー正則化を用いている。クラスタ内での相対的な重要度を表すために、 v_j の制約条件には従来のファジィクラスタリングとは異なり、 $\sum_{j=1}^m v_j = 1$ を用いるものとする。データ行列の要素ごとの近似を考えた際に、第 j 変量の誤差が大きい場合には v_j が小さくなり、誤差の小さい変量のメンバシップが相対的に大きくなる。すなわち、Fig. 1 のようなデータを 1 本の主成分ベクトルを用いて縮約する場合

には、他の変量との相関の小さい x_3 の近似誤差が大きくなることから、このような変量に小さな重みを与えて x_1 と x_2 の関連を強調した線形モデルの推定を行う。

3.2 局所的な主成分分析における定式化とアルゴリズム

つぎに、提案手法を局所的な主成分分析に拡張する。(6) 式では変量のメンバシップのみを用いているが、個体の分類を考慮したメンバシップ u_{ci} を同時に導入することで、以下の目的関数を定義する。

$$L_{pca_{uv}} = \sum_{c=1}^C \sum_{i=1}^n u_{ci} \sum_{j=1}^m v_{cj} \left(x_{ij} - \sum_{k=1}^p f_{cik} a_{ckj} - b_{cj} \right)^2 + \lambda_u \sum_{c=1}^C \sum_{i=1}^n u_{ci} \log u_{ci} + \lambda_v \sum_{c=1}^C \sum_{j=1}^m v_{cj} \log v_{cj} \quad (7)$$

ここで、 u_{ci} は個体の分類に寄与するパラメータであり、第 c クラスタに所属する標本データの選別を行う。一方、 v_{cj} は変量の選別に寄与するパラメータであり、第 c クラスタにおいて局所的な主成分を抽出する際に変量の重要度を決める。

クラスタごとに唯一の局所的線形モデルを推定するために、制約条件として、

$$F_c^\top U_c F_c = I \quad ; c=1, \dots, C \quad (8)$$

$$F_c^\top U_c \mathbf{1}_n = \mathbf{0} \quad ; c=1, \dots, C \quad (9)$$

$$\sum_{c=1}^C u_{ci} = 1 \quad ; i=1, \dots, n \quad (10)$$

$$\sum_{j=1}^m v_{cj} = 1 \quad ; c=1, \dots, C \quad (11)$$

および、 $A_c^\top A_c$ が対角行列となることを課すものとする。ただし、 $U_c = \text{diag}(u_{c1}, \dots, u_{cn})$ である。このモデルではクラスタ数が 1 のとき、すべての u_{ci} が 1 となり、前節での（単一の）主成分分析モデルに一致する。

個体の分類と局所的な主成分モデルの推定は、交互最適化に基づく繰り返しアルゴリズムを用いて行う。まず、 A_c 、 $\mathbf{b}_c$ および v_{cj} の更新則を導くために、(7) 式を以下のように置き換える。

$$L'_{pca_{uv}} = \sum_{c=1}^C \sum_{j=1}^m v_{cj} e_{cj}^2 + \lambda_u \sum_{c=1}^C \sum_{i=1}^n u_{ci} \log u_{ci} + \lambda_v \sum_{c=1}^C \sum_{j=1}^m v_{cj} \log v_{cj} \quad (12)$$

ただし、

$$e_{cj}^2 = (\mathbf{x}_j - F_c \tilde{\mathbf{a}}_{cj} - \mathbf{1}_n b_{cj})^\top U_c (\mathbf{x}_j - F_c \tilde{\mathbf{a}}_{cj} - \mathbf{1}_n b_{cj})$$

$$A_c = (\tilde{\mathbf{a}}_{c1}, \dots, \tilde{\mathbf{a}}_{cm})^\top$$

である。最適性の必要条件 $\partial L / \partial \tilde{\mathbf{a}}_{cj} = \mathbf{0}$ および $\partial L / \partial b_{cj} = 0$ から、更新則は以下のように与えられる。

$$\tilde{\mathbf{a}}_{cj} = (F_c^\top U_c F_c)^{-1} F_c^\top U_c (\mathbf{x}_j - \mathbf{1}_n b_{cj}) \quad (13)$$

$$b_{cj} = (\mathbf{1}_n^\top U_c \mathbf{1}_n)^{-1} \mathbf{1}_n^\top U_c (\mathbf{x}_j - F_c \tilde{\mathbf{a}}_{cj}) \quad (14)$$

また、 v_{cj} は $\partial L / \partial v_{cj} = 0$ および (11) 式の確率的な制約条件から、以下のように求められる。

$$v_{cj} = \frac{\exp\left\{-\frac{e_{cj}^2}{\lambda_v}\right\}}{\sum_{l=1}^m \exp\left\{-\frac{e_{cl}^2}{\lambda_v}\right\}} \quad (15)$$

同様に、 F_c および u_{ci} の更新則は、(7) 式を以下のように変形することで導かれる。

$$L''_{pca_{uv}} = \sum_{c=1}^C \sum_{i=1}^n u_{ci} d_{ci}^2 + \lambda_u \sum_{c=1}^C \sum_{i=1}^n u_{ci} \log u_{ci} + \lambda_v \sum_{c=1}^C \sum_{j=1}^m v_{cj} \log v_{cj} \quad (16)$$

ここで、

$$d_{ci}^2 = (\tilde{\mathbf{x}}_i - A_c \tilde{\mathbf{f}}_{ci} - \mathbf{b}_c)^\top \tilde{V}_c (\tilde{\mathbf{x}}_i - A_c \tilde{\mathbf{f}}_{ci} - \mathbf{b}_c)$$

である。最適性の必要条件 $\partial L / \partial \tilde{\mathbf{f}}_{ci} = \mathbf{0}$ から、

$$\tilde{\mathbf{f}}_{ci} = (A_c^\top \tilde{V}_c A_c)^{-1} A_c^\top \tilde{V}_c (\tilde{\mathbf{x}}_i - \mathbf{b}_c) \quad (17)$$

$$\tilde{V}_c = \text{diag}(v_{c1}, \dots, v_{cm}) \quad (18)$$

となる。また、 u_{ci} は $\partial L / \partial u_{ci} = 0$ および (10) 式の確率的な制約から、以下のように求められる。

$$u_{ci} = \frac{\exp\left\{-\frac{d_{ci}^2}{\lambda_u}\right\}}{\sum_{l=1}^C \exp\left\{-\frac{d_{li}^2}{\lambda_u}\right\}} \quad (19)$$

以上から、変量の重要度を表す重みを考慮した局所的な主成分分析のアルゴリズムは以下のように与えられる。

Algorithm: PCA considering relative responsibilities of variables

ステップ 1 クラスタごとに $U_c, \tilde{V}_c, A_c, \mathbf{b}_c, F_c$ を乱数により初期化し、(8)-(11) 式および $A_c^\top A_c$ が対角行列となる制約条件を満たすように基準化する。終了条件の閾値 ε を指定する。

ステップ 2 (13) 式を用いて A_c を更新し、 $A_c^\top A_c$ が対角行列となるように基準化する。

ステップ 3 (17) 式を用いて F_c を更新し、(8) 式および (9) 式を満たすように基準化する。

ステップ 4 (14) 式を用いて $\mathbf{b}_c$ を更新する。

ステップ 5 (15) 式を用いて $\tilde{V}_c$ を更新する。

ステップ 6 (19)式を用いて U_c を更新する.

ステップ 7 すべてのパラメータの変化量が ε 以下ならば終了. それ以外はステップ 2 へ戻る.

このように, 変量に関するメンバシップを用いることで, 不要な変数の影響を小さくしながら局所的線形モデルが推定される. しかし, (11)式のように変量についての和が1となるように基準化を行う場合には, 変数の数が増えるとメンバシップの絶対値が小さくなるという欠点がある. そのため, どの程度大きければ有意な変数であるかを定めることが困難となる. 同様に2種類のメンバシップと (11) 式に類する制約を用いる Pal らの Hybrid c -Means モデル [20] では, 可能性的な制約を用いることで改善が試みられている. そこで, 本論文でも可能性的なアプローチによる修正を行う.

3.3 段階的可能性アプローチを用いた変量の選択

本節では, 変量のメンバシップが絶対的な重要度を示すように, 和が1となる確率的な制約を取り払い, $v_{cj} \in [0, 1]$ のみを考慮する可能性的な制約 [16] を用いる. 可能性的な制約のもとでは, 変数の数が多い場合でも, 有意な変数には1に近いメンバシップが与えられるようになり, 変量の取捨選択の指標としての意味合いが強まる. しかしながら, 可能性的な制約を用いる際には, 分類の自由度が高まる反面, 初期分割への依存が強くなり, 大域的な最適解に収束しにくいという欠点もある [21]. 本論文では, 前節で得られる確率的な制約のもとでのメンバシップを絶対的な重要度に修正するという観点から, Masulli と Rovetta [17] により提案された段階的可能性アプローチを採用する.

変量に関するメンバシップの制約条件を, 以下のよう一般化して考える.

$$\Psi = \sum_{j=1}^m v_{cj}^{[\xi]} - 1 \quad (20)$$

ここで, $[\xi]$ はある区間 $[\underline{\xi}, \bar{\xi}]$ に含まれる任意の実数を表し, $\Psi = 0$ を満たすスカラー値 $\xi^* \in [\underline{\xi}, \bar{\xi}]$ が存在することを示すものとする. 簡単のために, 区間の境界値を単一のパラメータ $\alpha \in [0, 1]$ を用いて,

$$\underline{\xi} = \alpha, \quad \bar{\xi} = \frac{1}{\alpha} \quad (21)$$

と書くと, α を調節することで種々の制約条件を与えられる. たとえば, $\alpha = 1$ とすれば $[\xi] = [1, 1]$ となり, $\sum_{j=1}^m v_{cj} = 1$ のように確率的な制約となる. 一方, $\alpha \rightarrow 0$ とすれば $[\xi] = (0, \infty)$ となり, $\sum_{j=1}^m v_{cj}^0 > 1$ および $\sum_{j=1}^m v_{cj}^\infty < 1$ を与えることから, $v_{cj} \in [0, 1]$ のみを考慮する可能性的な制約に一致する. したがって, α を1から0へと徐々に減じることで, 確率的な制約条件から可能性的な制約条件へと段階的に移行することができる.

α を中間的な値とした場合, 制約条件は以下の二つの

不等式で与えられる.

$$\sum_{j=1}^m v_{cj}^\alpha \geq 1, \quad \sum_{j=1}^m v_{cj}^{1/\alpha} \leq 1 \quad (22)$$

この制約条件に従う領域は, $m=2$ とした場合, Fig. 2 に示す目のような形の領域となる. 領域は $\alpha=1$ のときに $v_{c1} + v_{c2} = 1$ の直線上にあり, α が小さくなるにつれて矢印の方向に拡大していく. $\alpha=0$ のときには正方形で囲まれた領域となり, 可能性的な制約条件に一致する.

Fig. 2 Bounds of the feasible regions for v_{cj} for different values of α

上記の段階的可能性アプローチを実装するため, Masulli と Rovetta [17] はエントロピー正則化クラスタリングにおける更新則を提案している. 主成分分析での変量選択に応用した場合, (23) 式ようになる.

$$v_{cj} = \frac{\phi_{cj}}{\kappa_j} \quad (23)$$

$$\phi_{cj} = \exp \left\{ -\frac{e_{cj}^2}{\lambda_v} \right\} \quad (24)$$

ただし, κ_j は以下にしたがって与える.

$$\kappa_j = \left(\sum_{l=1}^m \phi_{cl}^{1/\alpha} \right)^\alpha \quad \text{if } \sum_{l=1}^m \phi_{cl}^{1/\alpha} > 1 \quad (25)$$

$$\kappa_j = \left(\sum_{l=1}^m \phi_{cl}^\alpha \right)^{1/\alpha} \quad \text{if } \sum_{l=1}^m \phi_{cl}^\alpha < 1 \quad (26)$$

$$\kappa_j = 1 \quad \text{else} \quad (27)$$

以上の段階的可能性アプローチを採用することで, 確率的な制約のもとでのメンバシップを絶対的な重要度に修正する. 実装の際には, 3.2 のアルゴリズムにおける v_{cj} の更新式を (23) 式に変更し, α を1から十分小さい値になるまで減少させながらアルゴリズムを繰り返し実行する. すなわち, 確率的な制約で得られた結果を初期値とし, それらを段階的に可能性的な制約に緩和する. なお, 数値実験では v_{cj} のみを修正するために, u_{ci} は固定して段階的な緩和を行った.

Table 1 Artificial data set

sample	1	2	3	4	5	6	7	8	9	10	11	12
x_1	0.00	0.09	0.18	0.27	0.36	0.45	0.55	0.64	0.73	0.82	0.91	1.00
x_2	0.00	0.09	0.18	0.27	0.36	0.45	0.55	0.64	0.73	0.82	0.91	1.00
x_3	0.25	0.30	0.34	0.39	0.43	0.48	0.52	0.57	0.61	0.66	0.70	0.75
x_4	0.14	0.56	0.64	0.53	0.95	0.64	0.60	0.62	0.00	0.26	0.09	0.59
sample	13	14	15	16	17	18	19	20	21	22	23	24
x_1	0.20	0.41	0.37	0.95	0.58	0.32	0.90	0.40	0.25	0.21	0.84	0.17
x_2	0.75	0.70	0.66	0.61	0.57	0.52	0.48	0.43	0.39	0.34	0.30	0.25
x_3	0.25	0.30	0.34	0.39	0.43	0.48	0.52	0.57	0.61	0.66	0.70	0.75
x_4	0.00	0.09	0.18	0.27	0.36	0.45	0.55	0.64	0.73	0.82	0.91	1.00

Table 2 Memberships of variables and local principal component vectors (artificial data set)

variable	probabilistic constraint				possibilistic constraint			
	v_{cj}		a_c		v_{cj}		a_c	
	$c=1$	$c=2$	$c=1$	$c=2$	$c=1$	$c=2$	$c=1$	$c=2$
x_1	0.30	0.06	1.10	0.05	0.98	0.00	1.09	0.04
x_2	0.31	0.32	1.07	0.55	0.99	0.96	1.09	0.54
x_3	0.30	0.30	0.52	-0.50	0.95	0.98	0.54	-0.54
x_4	0.09	0.32	-0.22	-1.14	0.00	0.99	-0.24	-1.09

4. 数値実験

まず、提案手法における変量のメンバシップの役割を確かめるために人工的に作成したデータを用いた例を示した後に、ワインデータへの適用を通して潜在的な知識の発見法としての性能を評価する。ただし、段階的可能性アプローチでは $\alpha^{NEW} = \alpha^{OLD} - 0.1$ として制約の緩和を行った。

4.1 人工データを用いた数値実験

Table 1 に示す 4 次元 24 個の標本データからなる人工データを用いて実験を行った。データ集合は二つのグループからなる。標本 1–12 では x_1 から x_3 に一次元性があり、それらは 3 次元空間で直線上に散布されるのに対して、 x_4 は乱数で与えられたために他の変量との関連がない。一方、標本 13–24 では x_2 から x_4 に一次元性があるのに対して、 x_1 が関連のない乱数となっている。

まず、標本 1–12 のみを用いて、他の変量選択法との比較を行った。Jolliffe の主成分負荷量に着目した手法 [3] および Tanaka and Mori の拡張主成分分析の規準を利用した手法 [4] を適用したところ、いずれの手法でも x_1 から x_3 のいずれかと x_4 を重要な変量と判断した。このように、他の変量との関連がない無意味な変量が含まれる場合でも、冗長な変量を削除するという概念に基づく変量選択法ではそれらを尺度構成に必要な変量とみなしてしまうために、相関関係の抽出には適さない。また、単一のモデルにおける選択を前提とするために、局所的な主成分分析モデルにおいては利用できない。

そこで、24 個のデータすべてを用いて、データの分割と変量の重要度の推定を同時に考慮しながら局所的な主成分分析を施した。ただし、 $C=2$ 、 $p=1$ 、 $\lambda_u=0.01$ 、 $\lambda_v=0.1$ とした。データは標本 1–12 および 13–24 に対応する二つのクラスに分割され、クラスごとに局所的な主成分を抽出した。Table 2 に得られた変量のメンバシップおよび局所的な主成分ベクトルを示す。表中、左欄が確率的な制約を用いた際の結果を、右欄が段階的可能性アプローチにより可能性的な制約に移行した後の結果を示している。いずれの場合もクラスごとに乱数で与えられた変量に小さなメンバシップが割り当てられており、局所的な相関関係が強調された結果となっている。また、可能性的な制約に基づいて与えられたメンバシップでは変量の絶対的な重要度が $[0,1]$ の範囲で与えられており、選択の際の指標としての意味合いが強まっている。このように、提案手法では他の変量との関連が低く相関関係の抽出に不適な変量の影響を小さくするという概念に基づいて変量が選択されていることがわかる。

4.2 ワインデータを用いた数値実験

つぎに、UCI Machine Learning Repository [22] で公開されているワインデータを用いて実験を行った。ワインデータは 178 個の標本について 13 種類の変量を観測したデータ集合であり、おのおのの標本は三つのクラスのいずれかに属する。このデータ集合はしばしばパターン認識のベンチマークとして用いられるが、ほぼ線形分離が可能であり、三つのクラスがデータ空間内でそれぞ

Table 3 Factor loadings of component scores and memberships of variables ($C=1$)

variable	PCA	PCA _v	
	FL	FL	v_j
1	-0.313	-0.306	0.177
2	0.532	0.450	0.192
3	0.004	0.009	0.173
4	0.519	0.503	0.274
5	-0.308	-0.196	0.169
6	-0.856	-0.916	0.736
7	-0.917	-0.969	0.895
8	0.648	0.610	0.292
9	-0.680	-0.703	0.356
10	0.192	0.174	0.176
11	-0.644	-0.586	0.324
12	-0.816	-0.850	0.611
13	-0.622	-0.571	0.276

れ群をなして分布している。本研究では、教師なし分類の手法であるクラスタリングと主成分分析の融合により、13種類の変量の間に関連関係を抽出することを目的とする。ただし、各変量を平均が0、分散が1となるように基準化した後に分析を行った。

まず、単一の主成分分析モデルにより、データ集合全体の構造の把握を試みた。全体の大まかな特徴をつかむため、潜在変量の次元を $p=1$ とした。得られた潜在変量ともとの変量との相関係数に相当する因子負荷量 (FL) を Table 3 の第2, 3列に示す。表中、PCAが変量の選択を考慮しない主成分分析、PCA_vが提案法により可能性的な制約を用いて変量選択を行った場合の結果である。また、提案法により得られた変量のメンバシップを Table 3 の第4列に示す。通常の主成分分析の因子負荷量から、全体の特徴を表す潜在変量が第6, 7および12変量の3種類の観測値に関連が強いことがわかるが、提案手法ではそれらに大きなメンバシップが割り当てられており、相関関係をより強調した主成分分析が行われている。

つづいて、FCV法および提案手法により局所的な主成分分析を施した。ただし、各クラスに対応するデータ群の間に関連を調べるために、 $C=2$, $p=1$ とし、提案法では $\lambda_u=0.05$, $\lambda_v=50.0$ とおいた。おのおののクラスにおける各クラスの標本のメンバシップの平均値を Table 4 に示す。ただし、PCA_{uv}が提案法により得られた結果である。いずれの手法でも、第1クラスに所属する標本のほぼすべてが第1クラスに、第3クラスに所属する標本のほぼすべてが第2クラスに分類されたのに対して、第2クラスに所属する標本は、二つのクラスに分かれて割り当てられた。したがって、このデータ集合の局所的な線形構造を二つのクラスに分割して

Table 4 Average of memberships in each cluster

class label	FCV		PCA _{uv}	
	$c=1$	$c=2$	$c=1$	$c=2$
1	0.978	0.022	0.973	0.027
2	0.446	0.554	0.559	0.441
3	0.034	0.966	0.021	0.979

Table 5 Cluster centers and fuzzy factor loadings of component scores derived by FCV algorithm ($C=2$)

variable	b_c		FFL	
	$c=1$	$c=2$	$c=1$	$c=2$
1	0.241	-0.252	0.871	0.619
2	-0.298	0.312	-0.048	0.698
3	0.275	-0.288	-0.049	0.590
4	-0.239	0.250	-0.698	0.451
5	0.178	-0.187	0.514	0.163
6	0.512	-0.536	0.785	-0.667
7	0.608	-0.637	0.768	-0.817
8	-0.224	0.234	-0.661	0.603
9	0.410	-0.430	0.442	-0.484
10	-0.149	0.156	0.755	0.776
11	0.410	-0.429	0.019	-0.835
12	0.547	-0.573	0.414	-0.855
13	0.496	-0.519	0.872	0.334

分析する際には、第1クラスと第2クラスの間を表すモデル (第1クラス) と、第2クラスと第3クラスの間を表すモデル (第2クラス) の2種類を推定すればよいことがわかる。FCV法で得られたクラスタ中心 (b_c) およびファジィ因子負荷量 (FFL) を Table 5 に示す。また、提案法により得られたクラスタ中心、変量のメンバシップおよびファジィ因子負荷量を Table 6 に示す。メンバシップが太字で示された変量が局所的な主成分分析モデルの推定で重要とみなされた変量を表している。FCV法の結果と比較して、選択された変量の因子負荷量の絶対値が増加していることから、潜在的な相関関係を強調して分析が行われていることがわかる。一方、第1クラスタの第2,12変量や第2クラスタの第13変量のように、因子負荷量の絶対値が小さいにもかかわらず大きなメンバシップを持つ変量もあるが、これらはクラスタ中心の値に偏りが大きく、標本の分類に用いられたためにクラスタ内での分散が小さくなった変量であることを示している。すなわち、標本の層別を解釈するうえで重要な役割を担う変量であるといえる。このように、変量ごとの近似誤差をクラスタリング基準として推定されるメンバシップには、潜在的な相関関係の抽出における重要度とともに標本の層別における重要度を表す働きがある。

Table 6 Cluster centers, memberships of variables and fuzzy factor loadings of component scores derived by proposed method ($C = 2$)

variable	b_c		FFL		v_{cj}	
	$c = 1$	$c = 2$	$c = 1$	$c = 2$	$c = 1$	$c = 2$
1	0.199	-0.245	0.896	0.683	0.633	0.575
2	-0.474	0.583	0.213	0.377	0.572	0.184
3	-0.010	0.012	0.285	0.165	0.111	0.326
4	-0.346	0.425	-0.469	0.101	0.209	0.367
5	0.120	-0.147	0.433	0.095	0.174	0.270
6	0.387	-0.475	0.875	-0.857	0.716	0.678
7	0.513	-0.630	0.867	-0.919	0.770	0.813
8	-0.283	0.348	-0.490	0.476	0.350	0.247
9	0.295	-0.363	0.534	-0.668	0.333	0.407
10	-0.183	0.225	0.811	0.727	0.665	0.354
11	0.561	-0.689	-0.179	-0.713	0.439	0.524
12	0.511	-0.628	0.454	-0.921	0.510	0.791
13	0.428	-0.526	0.873	0.357	0.572	0.716

5. おわりに

本論文では、主成分分析を用いた知識発見への新たなアプローチとして、相関関係を強調した変量の選択を行う主成分分析モデルをファジィクラスタリングとの融合により提案した。ファジィクラスタリングにおいて個体の分類に用いられるメンバシップを変量の選択に応用することで、不要な変量の影響を小さくしながら線形モデルが推定される。変量のメンバシップの制約には確率的な制約を課すこともできるが、段階的可能性アプローチにより可能性的な制約に移行することにより、変量の重要度に関する絶対的な指標として用いることも可能である。また、個体のメンバシップを同時に考慮することにより、提案手法は局所的な主成分分析モデルへと拡張が可能であり、データの層別と相関関係の抽出を同時に考慮したデータマイニングツールとしての利用が期待される。ただし、変量の選択を考慮した場合、潜在変量の次元が異なれば選ばれる変量も異なる。したがって、分散共分散行列の最大固有値から順に対応した固有ベクトルを採用する通常の主成分分析のように、次元の異なるモデルを比較する寄与率のような指標が求められない。さらに、他のファジィクラスタリング手法と同様にメンバシップのファジィ度やクラスタ数は分析者が定める必要があり、それらを定める際の指標の開発が今後の課題である。

参考文献

- [1] 上田: データマイニング入門講座 1~12(連載); 企業診断, 1~12月号, 同友館 (2000)
- [2] VASpca Web Page;
<http://mo161.soci.ous.ac.jp/vaspca/>
- [3] I. T. Jolliffe: Discarding variables in a principal component analysis. I. Artificial data; *Appl. Statist.*, Vol. 21, pp. 160–173 (1972)
- [4] Y. Tanaka and Y. Mori: Principal component analysis based on a subset of variables: Variable selection and sensitivity analysis; *American Journal of Mathematics and Management Sciences*, Vol. 17, No. 1&2, pp. 61–89 (1997)
- [5] J. C. Bezdek: *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press (1981)
- [6] J. C. Bezdek, C. Coray, R. Gunderson and J. Watson: Detection and characterization of cluster substructure 2. fuzzy c -varieties and convex combinations thereof; *SIAM J. Appl. Math.*, Vol. 40, No. 2, pp. 358–372 (1981)
- [7] G. E. Hinton, P. Dayan and M. Revow: Modeling the manifolds of images of handwritten digits; *IEEE Trans. Neural Networks*, Vol. 8, No. 1, pp. 65–74 (1997)
- [8] M. E. Tipping and C. M. Bishop: Mixtures of probabilistic principal component analysers; *Neural Computation*, Vol. 11, No. 2, pp. 443–482 (1999)
- [9] Y. Mori, K. Honda, A. Kanda and H. Ichihashi: A unified view of probabilistic PCA and regularized linear fuzzy clustering; *IJCNN 2003 Conference Proceedings*, pp. 541–546 (2003)
- [10] K. Honda and H. Ichihashi: Regularized linear fuzzy clustering and probabilistic PCA mixture models; *IEEE Trans. Fuzzy Syst.*, Vol. 13, No. 4 (2005)
- [11] P. Whittle: On principal components and least square methods of factor analysis; *Skand. Akt.*, Vol. 36, pp. 223–239 (1952)
- [12] K. Honda and H. Ichihashi: Linear fuzzy clustering techniques with missing values and their application to local principal component analysis; *IEEE Trans.*

- Fuzzy Syst.*, Vol. 12, No. 2, pp. 183–193 (2004)
- [13] K. Honda and H. Ichihashi: Component-wise robust linear fuzzy clustering for collaborative filtering; *Int. J. Approximate Reasoning*, Vol. 37, No. 2, pp. 127–144 (2004)
- [14] A. Gifi: *Nonlinear Multivariate Analysis*, Wiley (1990)
- [15] 土屋: 項目分類のための数量化法; 行動計量学, Vol. 22, No. 2, pp. 95–109 (1995)
- [16] R. Krishnapuram and J. M. Keller: A possibilistic approach to clustering; *IEEE Trans. Fuzzy Syst.*, Vol. 1, pp. 98–110 (1993)
- [17] F. Masulli and S. Rovetta: The graded possibilistic clustering model; *IJCNN 2003 Conference Proceedings*, pp. 791–796 (2003)
- [18] 宮本, 馬屋原, 向殿: ファジィc-平均法とエントロピー正則化法におけるファジィ分類関数; 日本ファジィ学会誌, Vol. 10, No. 3, pp. 156–164 (1998)
- [19] Y. Yabuuchi and J. Watada: Fuzzy principal component analysis and its application; *Biomedical Fuzzy and Human Sciences*, Vol. 3, No. 1, pp. 83–92 (1997)
- [20] N. R. Pal, K. Pal, J. M. Keller and J. C. Bezdek: A new hybrid c-means clustering model; *Proc. of 2004 IEEE International Conference on Fuzzy Systems*, #0026-1379 (2004)
- [21] 渋谷, 宮本, 高田, 馬屋原: ファジィc-平均クラスタリングにおける制約条件と正則化および可能性クラスタリング; 日本ファジィ学会誌, Vol. 13, No. 6, pp. 707–715 (2001)
- [22] P. M. Murphy and D. W. Aha: UCI Repository of machine learning databases; <http://www.ics.uci.edu/~mlearn/MLRepository.html> Irvine, CA: University of California, Department of Information and Computer Science (1994)