# 相関係数による変数のクラスタリングを用いた部分空間法の提案

1X11C075-6　田原將志
指導教員　後藤正幸

## 1　研究背景と目的

近年，データベース上に蓄積されるデータの高次元化が進んでおり，高次元データを分析する手法の重要性が増している．このデータを分類する手法の 1 つに部分空間法 [1] が挙げられる．部分空間法は，学習データをカテゴリに分割し，カテゴリごとの主成分分析により部分空間を構成する正規直交基底を求め，その部分空間と入力データの距離を計算してカテゴリ判定を行う手法である．一般に，文書データでは各単語の文書中の出現回数を一つの変数とみなすため，多くの変数を含む高次元なデータ構造となり，相関の低い変数の組が多く含まれる傾向がある．このようなデータに対し部分空間法を適用すると，相関の低い変数の数が膨大であることから，その影響が相対的に大きくなり，分類に有効な部分空間を構成する正規直交基底の推定精度を劣化させる可能性がある．

そこで本研究では高次元なデータを対象とし，部分空間法による分類精度向上のため，各カテゴリで相関の低いと思われる変数を学習データから削除する方法を提案する．これにより，相関の低い変数の影響を受けずに部分空間を構成する正規直交基底を求めることが可能になる．具体的には，カテゴリごとに変数の相関係数を算出し，相関関係のある変数群と，どのような変数とも相関を持たない変数群が存在することに着目する．そこで相関係数を用いて変数間の距離尺度を定義し，階層的クラスタリング手法の 1 つであるウォード法を用いて互いに相関の高い変数同士を同一のサブグループへクラスタリングする．このサブグループごとに主成分分析を行うことで，相関の低い変数の影響を受けずに部分空間を構成する正規直交基底を求めることができる．

提案手法を新聞記事データを用いた文書分類問題に適応し，その有効性を検証する．

## 2　準備

### 2.1　主成分分析

主成分分析は，学習データに対し分散が最大となる方向への線形射影を順次求める手法である．学習データの次元数を $d$，サンプル数を $N$ とした時，変数 $x_a$ と $x_b$ の相関係数 $r_{a,b}$ は，

$$r_{a,b} = \frac{\sum_{i=1}^{N}(x_{a,i}-\bar{x}_a)(x_{b,i}-\bar{x}_b)}{\sqrt{\sum_{i=1}^{N}(x_{a,i}-\bar{x}_a)^2}\sqrt{\sum_{i=1}^{N}(x_{b,i}-\bar{x}_b)^2}} \tag{1}$$

と表され，学習データの相関係数行列 $\boldsymbol{R}$ は，

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{1,2} & \dots & r_{1,d} \\ r_{2,1} & 1 & \dots & r_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ r_{d,1} & r_{d,2} & \dots & 1 \end{pmatrix} \tag{2}$$

と定義される．この相関係数行列 $R$ の固有ベクトル $\{\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_d\}$ が学習データの主成分ベクトルとなり，その固有値 $\lambda_1, \lambda_2, ..., \lambda_d$ を用いて，主成分 $\boldsymbol{u}_i$ 寄与率を，

$$\Lambda_i = \frac{\lambda_i}{\sum_{j=1}^{d}\lambda_j} \tag{3}$$

と表す．

### 2.2　部分空間法（CLAFIC 法）

部分空間法は，カテゴリごとに部分空間を構成する基底ベクトルを学習データから求め，入力データを各カテゴリの部分空間に射影して分類する手法である．データ数 $N$ の学習データの集合を $\mathcal{X} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ とする．ここで，$\boldsymbol{x}_i$ は $d$ 次元特徴ベクトルで，$y_i \in \{1, 2, ..., K\}$ は $\boldsymbol{x}_i$ が属するカテゴリとする．カテゴリごとの部分空間を $\boldsymbol{P}_1, ..., \boldsymbol{P}_K$ とし，カテゴリ $k$ の部分空間 $\boldsymbol{P}_k$ を構成する正規直交基底ベクトルを $\{\boldsymbol{u}_{k1}, ..., \boldsymbol{u}_{kd_k}\}$ と表す．ただし $d_k$ は部分空間 $\boldsymbol{P}_k$ の次元数である．まずカテゴリごとのデータ集合 $\mathcal{X}^k (k = 1, 2, ..., K)$ を作成し，個別に主成分分析を行うことで部分空間を構成する基底ベクトルの候補 $\{\boldsymbol{u}_{k1}, \boldsymbol{u}_{k2}, ..., \boldsymbol{u}_{kd}\}$ を求め，これを並べた行列を $\boldsymbol{U}_k$ とする．次に固有値から求められる累積寄与率，

$$a(d_k) = \frac{\sum_{j=1}^{d_k}\lambda_{kj}}{\sum_{j=1}^{d}\lambda_{kj}} \tag{4}$$

と任意のパラメータ $\kappa$ を用いて，

$$a(d_k - 1) < \kappa < a(d_k) \tag{5}$$

を満たす $d_k$ を部分空間 $\boldsymbol{P}_k$ の次元数とし，カテゴリ $k$ の部分空間を構成する基底ベクトルを $\{\boldsymbol{u}_{k1}, ..., \boldsymbol{u}_{kd_k}\}$ とし，これを並べた行列を $\boldsymbol{U}_{\boldsymbol{P}_k} \in \mathcal{R}^{d \times d_k}$ とする．データ $\boldsymbol{x}$ の部分空間 $\boldsymbol{P}_k$ への射影長を $\sum_{j=1}^{d_k}\boldsymbol{x}^T\boldsymbol{u}_{kj}\boldsymbol{u}_{kj}^T\boldsymbol{x}$ とし，これを最大とするカテゴリへ分類する．

## 3　提案手法

### 3.1　概要

一般的な部分空間法では，カテゴリごとに全ての変数を一括りにして部分空間の正規直交基底ベクトルを求めるため，多くの変数を含む高次元データでは，互いに相関の低い変数の組が多く含まれることから，それらの影響が相対的に大きくなり，分類に有効な部分空間を構成する正規直交基底の推定精度を劣化させる可能性がある．そこで本研究では，相関係数を用いて変数間の距離尺度を定義し，階層的クラスタリング手法の 1 つであるウォード法を用いて相関の高い変数を同一のサブグループへクラスタリングを行い，サブグループごとに主成分分析を行うことで，相関の低い変数の影響を抑え部分空間を構成することができ，分類精度の向上が見込まれる．提案手法の概要を図 1 に示す．

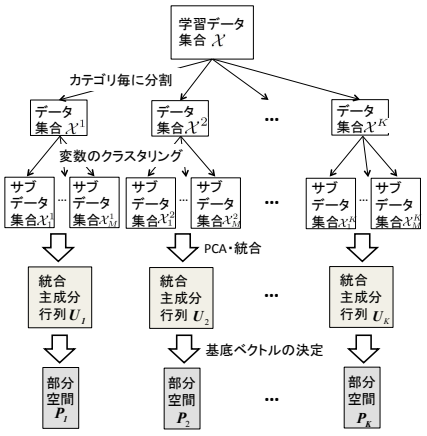


**図 1．提案手法の概要**

### 3.2 変数のクラスタリングと基底ベクトルの算出法

カテゴリごとのデータ集合 $\mathcal{X}^k(k=1,2,...,K)$ に対して，式 (4) で表される相関係数を計算したものを $r^k_{a,b}$ とし，カテゴリ $k$ における変数 $x_a$ と $x_b$ の距離を表す尺度 $d^k_{a,b}$ を，

$$d^k_{a,b} = \frac{1}{|r^k_{a,b}|} \tag{6}$$

と定義し，ウォード法を用いて変数を $M$ 個のサブグループへとクラスタリングを行う．カテゴリごとのデータ集合 $\mathcal{X}^k$ をサブグループごとに分割し，サブデータ集合 $\mathcal{X}^k_m(m=1,2,...,M)$ を作成し，個別に主成分分析を行いサブグループごとの主成分の係数行列 $\boldsymbol{U}_{k,m}$ を求める．サブグループごとの主成分の係数行列 $\boldsymbol{U}_{k,m}$ を式 (7) のように統合し $\boldsymbol{U}^c_k \in \mathcal{R}^{d\times d}$ を作成する．$\boldsymbol{U}^c_k$ を元の変数の順に行ベクトルをソートし，部分空間 $\boldsymbol{P}_k$ を構成する基底ベクトルの候補 $\{\boldsymbol{u}_{k1}, \boldsymbol{u}_{k2}, ..., \boldsymbol{u}_{kd}\}$ を得る．これを並べた行列を統合主成分行列 $\boldsymbol{U}_k \in \mathcal{R}^{d\times d}$ とする．

$$\boldsymbol{U}^c_k = \begin{pmatrix} \boldsymbol{U}_{k,1} & (0) & \dots & (0) \\ (0) & \boldsymbol{U}_{k,2} & \dots & (0) \\ \vdots & \vdots & \ddots & \vdots \\ (0) & (0) & \dots & \boldsymbol{U}_{k,M} \end{pmatrix} \tag{7}$$

本手法ではサブグループごとに主成分分析を行うため，式 (4) で表される累積寄与率を用いて $\boldsymbol{U}_k$ の次元削減を行い，部分空間 $\boldsymbol{P}_k$ を構成する基底ベクトルを求めることはできない．そこで学習データを主成分 $\boldsymbol{u}_i$ に射影した後の分散値 $V_i = \boldsymbol{u}_i^T Var\{\bar{\boldsymbol{X}}\}\boldsymbol{u}_i$ が，一般的な主成分分析により得られる固有値 $\lambda_i$ と類似した意味を持つことに着目する．まず主成分 $\boldsymbol{u}_i$ へ射影した後の分散値 $V_i$ の降順に $\boldsymbol{U}_k$ の列ベクトルをソートし $\boldsymbol{U'}_k = (\boldsymbol{u'}_{k1}, \boldsymbol{u'}_{k2}, ..., \boldsymbol{u'}_{kd})$ を得る．次に射影後の分散値の累積値の比率，

$$b(d_k) = \frac{\sum_{j=1}^{d_k} V_{kj}}{\sum_{j=1}^{d} V_{kj}} \tag{8}$$

と任意のパラメータ $\kappa$ を用いて，

$$b(d_k - 1) < \kappa < b(d_k) \tag{9}$$

を満たす $d_k$ を部分空間 $\boldsymbol{P}_k$ の次元数とし，部分空間 $\boldsymbol{P}_k$ を構成する基底ベクトルを $\{\boldsymbol{u'}_{k1}, ..., \boldsymbol{u'}_{kd_k}\}$ とし，これを並べた行列を $\boldsymbol{U}_{P_k} \in \mathcal{R}^{d\times d_k}$ とする．

### 3.3 提案アルゴリズム

以下に提案手法の学習アルゴリズムを示す．

**Step1)** 学習データ集合 $\mathcal{X}$ をカテゴリごとに分割し，カテゴリごとのデータ集合 $\mathcal{X}^k(k=1,2,...,K)$ を作成する．

**Step2)** カテゴリごとのデータ集合 $\mathcal{X}^k$ に対して相関係数行列を計算する．式 (6) より求められる変数間の距離尺度を用いて，ウォード法により変数のクラスタリングを行い，カテゴリごとのサブデータ集合 $\mathcal{X}^k_m(m=1,2,...,M)$ を作成する．

**Step3)** カテゴリごとのサブデータ集合 $\mathcal{X}^k_m$ に対して個別に主成分分析を行い，主成分の係数行列 $\boldsymbol{U}_{k,m}$ を求める．

**Step4)** $\boldsymbol{U}_{k,m}$ を式 (7) により統合し $\boldsymbol{U}^c_k$ を作成．$\boldsymbol{U}^c_k$ を元の変数の順に行をソートし，統合主成分行列 $\boldsymbol{U}_k = (\boldsymbol{u}_{k1}, \boldsymbol{u}_{k2}, ..., \boldsymbol{u}_{kd})$ を作成する．

**Step5)** 統合主成分行列 $\boldsymbol{U}_k$ の主成分ベクトル $\boldsymbol{u}_{ki}(i=1,2,...,d)$ へ学習データ集合を射影した後の分散値 $V_{ki} = \boldsymbol{u}_i^T Var\{\bar{\boldsymbol{X}}\}\boldsymbol{u}_i$ を計算し，$V_{ki}$ の降順に $\boldsymbol{U}_k$ の列ベクトルをソートし $\boldsymbol{U'}_k = (\boldsymbol{u'}_{k1}, \boldsymbol{u'}_{k2}, ..., \boldsymbol{u'}_{kd})$ を得る．

**Step6)** 式 (9) を満たす $d_i$ を計算し，カテゴリ $k$ の部分空間を構成する基底ベクトル $\boldsymbol{P}_k = (\boldsymbol{u'}_{k1}, ..., \boldsymbol{u'}_{kd_k})$ とし，これを並べた行列を $\boldsymbol{U}_{P_k}$ とする．

□

## 4 実験

### 4.1 実験条件

提案手法の有効性を示すため新聞記事を用いた分類実験を行い，分類正解率の評価を行った．分類正解率は式 (10) により計算する．

$$\text{分類正解率} = \frac{\text{正しく分類されたテストデータ数}}{\text{テストデータ数}} \tag{10}$$

実験では 2010 年の毎日新聞記事から 9 カテゴリ（社説，国際，経済，家庭，文化，読書，芸能，スポーツ，社会）を使用した．学習データは各カテゴリ 100 件とし，テストデータは各カテゴリ 100 件とした．また予備実験の結果より累積寄与率 $\kappa$，サブグループ数 $M$ を表 1 の通り設定した．

**表 1. データセットの概要とパラメータ**

| 次元数 | カテゴリ数 | 学習データ数 | テストデータ数 | $\kappa$ | $M$ |
|---|---|---|---|---|---|
| 2172 | 9 | 900 | 900 | 0.99 | 3 |

比較手法として変数のクラスタリングを行わない，一般的な部分空間法を用いた．

### 4.2 結果と考察

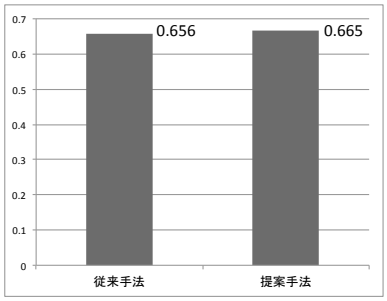


**図 2. 実験結果**

図 2 より変数を分割するサブデータ数 3 において，提案手法の分類精度が従来手法よりも優れていることが分かる．これは変数のクラスタリングを行い，相関の強い変数のみで主成分分析を行ったことで，相関の低い変数の影響を抑え部分空間を構成することができたためであると考えられる．

## 5 まとめと今後の課題

本研究では，相関係数による変数のクラスタリングを用いた部分空間法を提案し，評価実験により分類精度の向上が確認された．今後の課題として，距離尺度やクラスタリング手法の改善，変数をクラスタリングするサブグループ数 $M$ の自動決定法の考案や，変数が複数のクラスタに所属するモデルの考案などが挙げられる．

## 参考文献

[1] Satoshi Watanabe, “Knowing and Guessing:A Quantitative Study of Infrence and Information,” *John Wiley & Sons,Inc,New York*, 1969.