直交空間での教師なし類別についての基礎研究

瀬　良　豊　士*

A Study of Unsupervised Clustering in an Orthogonal System

Toyoshi SERA

Abstract

It is known that the Karhunen-Loéve system can be applied to unsupervised clustering in an orthogonal system. And some results of computer-simulation are obtained with the method for unsupervised clustering by the K-L system. But, in the case of the practical patterns, the results are not so much given. So, first the principle of the method for unsupervised clustering by the K-L system are explained here. Secondly, English character, A and B are used as the practical examples and computer-simulation by the method described above is carried out.

1. 緒　言

パターン認識の問題は，対象とするものについての予備知識があるかないかによって，次の2つに分けられる．その1つは，まえもって知られているクラス集合のいずれかに与えられた未知パターンを決定するパターン認識の問題である．他の1つは，多数の代表パターン集合から共通なもの同志を同一クラスとしてまとめクラスの集合を形成する **clustering** の問題，すなわち cognition の問題である．ここでは，未知パターンベクトルを直交成分に展開する Karhunen-Loéve 法を用いて，分布の未知なパターン集合からクラスの集合を形成する clustering の一方法を考察する．この問題は，すでに Watanabe[1], Fukunaga[2], 富田[3] 等によって取り扱われているが，実際のパターンを用いた実験結果は余り得られていない．そこで．ここでは，まず KL-法を用いた教師なしの類別法が，距離の概念を用いた類別法であることを示す．つぎに，実際の未知パターン集合として，手書きの文字A，Bを用いて行なったシミュレーションの実験結果を示す．

2. Karhunen-Loéve 直交系

Karhunen-Loéve 展開は与えられたパターンから特徴を抽出する最も有力な方法である．ここでは，後にこの方法を教師なしの類別問題に適用するため，まず Karhunen-Loéve 展開を導入する．この展開方法は確率ベクトルを直交ベクトル成分の線形結合で表示する展開方法である．いま，分布が未知なる正規化されたパターン $X_{(i)}^{k}$ の集合 S を

$$S=\{X_{(i)}^{k} \mid k=1,2,\cdots,\ M\ i=1,2,\cdots\cdots,\ N\} \quad \cdots\cdots\cdots\cdots(1)$$

ただし，$X_i^{(k)}$ ：m 次元パターンベクトル，

$$X_i^{(k)T}=(x_{i1}^{(k)},\ \cdots\cdots,\ x_{im}^{(k)}).$$

M：パターンクラスの数．

N：任意に抽出されたサンプルパターンの数．

とすると，パターンベクトル $X_k^{(i)}$ は

$$X_i^{(k)}=\sum_{\ell=1}^{m} a_{i\ell}^{(k)}\xi_\ell \quad \cdots\cdots\cdots(2)$$

ただし，$\{\xi_\ell \mid \ell=1,2,\cdots\cdots m\}$：正規直交ベクトル．(3)

$$(\xi^i\cdot\xi_j)=\delta_{ij}\cdot$$

$$\{a_{i\ell}^{(k)} \mid k=1,2,\cdots\cdots M\quad i=1.2.\cdots\cdots N \quad \ell=1,2\cdots\cdots\cdots m\}$$

：展開係数

$$a_{i\ell}^{(k)}=(X_i^{(k)}\cdot\xi_\ell) \quad (4)$$

と展開されることがすでに明らかになっている．[1),4)] しかし，実際に $X_i^{(k)}$ を(2)式の様に展開するためには(3)の要素を求める必要がある．この要素は各クラスの生

* 工業短期大学部電気工学科

起確率を P_k とすれば，つぎの自己相関行列の固有ベクトルとして求められることが分っている．

$$G=\sum_{k=1}^{M} P_k \cdot E\left(X_i^{(k)} \cdot X_i^{(k)T}\right) \qquad (5)$$

$$\sum_{k=1}^{M} P_k = 1$$

$E(X_i^{(k)} \cdot X_i^{(k)T})$：クラスkに関する自己相関行列．

したがって．この対称行列Gの固有値 λ_l の集合を

$$\{\lambda_l \mid l=1,2,\cdots\cdots,m\} \qquad (6)$$

とすれば，(3)の要素は

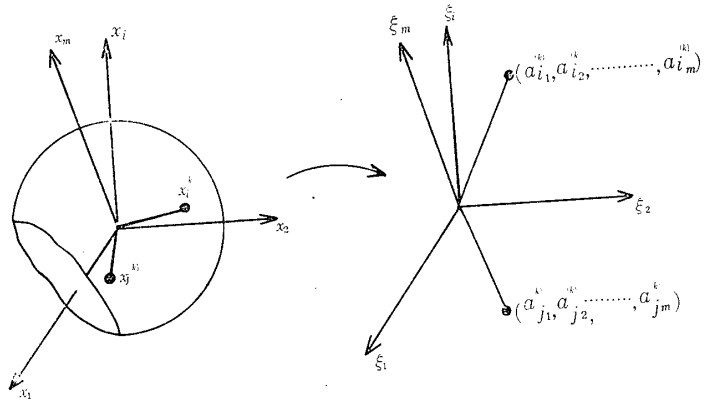
$$G\,\xi_l=\lambda_l\,\xi_l \qquad (7)$$

から求められる．この求められた $\{\xi_l\}$ を Karhunen-Loéve 直交系，略して K-L という．さらに(2)式を(6)式に代入すると．

$$\lambda_l=\sum_{k=1}^{m} P_k \cdot Var(a_{il}^{(k)}) \qquad (8)$$

$Var(a_{il}^{(k)})$：$a_{il}^{(k)}$ の分散

と求められる．この結果，(4)式と(8)式から λ_l は直交ベクトル ξ_l のパターンを表示する重要度を表わしていることが分かる．このことは $\{\xi_l\}$ の要素を座標軸とする空間にパターンベクトルが写像された状態を考えることによってより明らかになる．この様子をFig. 1に示す．



Ftg. 1 Mapping two pattern vectos into the K-L orthogonal space

3. 教師なしの類別

2つの未知分布から抽出されたパターン集合を同じクラス同志に分ける2類別問題を考える．したがってパターン集合は，

$$S=\{X_i^{(k)} \mid k=1,2\ i=1,2,\cdots\cdots N\} \cdots(9)$$

となる．この集合を2つのクラス Class 1，Class 2 に教師なしで類別するためには，教師の代りとなる1つの criterion を決めておく必要がある．そこで，クラス1とクラス2の生起確率が同じであるとすれば，各クラスの自己相関行列は

$$G_1=\frac{1}{N}\sum_{l=1}^{N} X_i^{(1)}\,X_i^{(1)T} \qquad (10)$$

$$G_2=\frac{1}{N}\sum_{l=1}^{N} X_i^{(2)}\,X_i^{(2)T} \qquad (11)$$

$$X_i^{(1)}=\sum_{l=1}^{m} a_{il}^{(1)}\xi_l\,, \qquad X_i^{(2)}=\sum_{l=1}^{m} a_{il}^{(2)}\xi_l \qquad (12)$$

となり，これらの式から G_1 行列と G_2 行列の差の行列 G_0 のトレースは，

$$t_r(G_0)=t_r(G_1-G_2)$$

$$=t_r\left(\frac{1}{N}\left(\sum_{i=1}^{N} X_i^{(1)} \cdot X_i^{(1)T} - \sum_{i=1}^{N} X_i^{(2)} \cdot X_i\right)\right)$$

$$=t_r\begin{pmatrix} \frac{1}{N}\left(\sum_{i=1}^{N}\left(a_{i1}^{(1)}\right)^2-\sum_{i=1}^{N}\left(a_{i1}^{(2)}\right)^2\right) & & 0 \\ & \frac{1}{N}\left(\sum_{i=1}^{N}\left(a_{i2}^{(1)}\right)^3-\sum_{i=1}^{N}\left(a_{i2}^{(2)}\right)^2\right) & \\ 0 & & \frac{1}{N}\left(\sum_{i=1}^{N}\left(a_{im}^{(2)}\right)^2-\sum_{i=1}^{N}\left(a_{im}^{(2)}\right)^2\right) \end{pmatrix}$$

$$=t_r\begin{pmatrix} \lambda_1^{(1)}-\lambda_1^{(2)} & & 0 \\ & \lambda_2^{(1)}-\lambda_2^{(2)} & \\ 0 & & \lambda_m^{(1)}-\lambda_m^{(2)} \end{pmatrix} \qquad (13)$$

と求められる．これから行列 G_0 の対角要素は，$\{\xi_l\}$ を座標とした空間を考えた場合，各座標軸成分での2つのパターンクラス Class 1 と Class 2 の特徴差を確率平均の意味で表わしていることが分かる．このことはFig. 1を参照すると良く分かる．したがって，**距離の概念**を導入して，つぎなる教師なしの類別の**Griterion C**

$$C=\sum_{l=1}^{N}(\lambda_l^{(1)}-\lambda_l^{(2)})^2=\sum_{l=1}^{N}(\lambda_l^{(0)})^2 \qquad (14)$$

$\lambda_l^{(1)}-\lambda_l^{(2)}=\lambda_l^{(0)}$ ：G_0の固有値

を採用できる．この結果Cを最大にするようにパターンを Class 1と Class 2に分割すれば最適な類別が行なわれることが分かる．さらに(14)式は

$$C=\sum_{l=1}^{N}(\lambda_l^{(0)})^2=t_r(G_0)^2 \qquad (15)$$

t_r：$(G_0)^2$ のトレース

となる．

4. サンプルパターン交換のアルゴリズム

教師なしの類別に必要な criterion C は(14)式として求められたが，実際にこのCを最適にもっていくためにはサンプルパターンの交換を繰り返し行なう必要がある．いま，Class 1のパターン $X_r^{(1)}$ と Class 2のパターン $X_t^{(2)}$ を交換したとき，行列G_0の変化を考えると．

$$\triangle G=\frac{2}{N}(X_t^{(2)}\cdot X_t^{(2)T}-X_r^{(1)}\cdot X_r^{(1)T}) \qquad (16)$$

となる．したがって交換後のG_0行列は

$$G=G_0+\triangle G \qquad (17)$$

となる．このとき Giterion C の変化を $\triangle C$ とすると

$$\begin{aligned}\triangle C&=\sum_{l=1}^{N}\left((\lambda_l^{(0)}+\triangle\lambda_l^{(0)})^2-(\lambda_l^{(0)})^2\right)\\&=t_r(G_0+\triangle G)^2-t_r(G_0)^2\\&=2\,t_r(G_0\cdot\triangle G)+t_r(\triangle G)^2 \qquad (18)\end{aligned}$$

となる．これから，正しく類別するには，すなわちCを最大にもっていくには，$\triangle C>0$ のときのみにClass 1 と Class 2 のパターンを交換することによって行なわれる．$t_r(\triangle G)^2$ が常に正の値であることを考えるなら

$$\triangle C=2\,t_r(G_0\cdot\triangle G) \qquad (19)$$

とすることが出来る．以上より交換を行なうアルゴリズムはつぎのようになる．

1）$(X_r^{(1)}\cdot X_r^{(1)T})$ を r に固定し，t を変化させ $\triangle G$ を求める．

2）$\triangle C>0$ なら，この値が最大なる Class 2 のパターン $X_t^{(2)}$ を交換をする．

3）$r=r+1$ として，$r=N$ になるまで(1)，(2)を繰り返す．

4）$\triangle C$ がもはや変化しなくなったらストップする．

5）$r=N$ でも $\triangle C$ が変化するなら初めから繰り返す．

5. シミュレーション実験

集合Sのパターンとして手書文字A，Bから再構成された9次元の低次元パターンベクトルを採用した．[5,6] パターン数は全部で30個とし各クラスから任意に15個抽出した．したがって，シミュレーション実験が，

$$S=\left\{X_i^{(k)}\;\middle|\;k=1,2\quad i=1,2,\cdots\cdots 15\right\}\cdots(20)$$

の条件のもとで行なわれた．計算は$\triangle C$が0.005以下になったときに停止するようにした．この結果，9回のパターンの交換によって類別が終了した．類別結果は30個の内2個のパターンが誤まって類別されたのみである．Fig. 2に criterion C の変化の様子を示す．Table 1，2には，初期の状態における **Class 1** と **Class 2** のパターン集合，Table 3，4には計算終了状態における **Class 1** と **Class 2** のパターン集合を示す．Fig. 3に，シミュレーション実験を行った計算機プログラムのフローチャートを示す．なお，このシュレーション実験は TOSBAC-3400 によって行った．

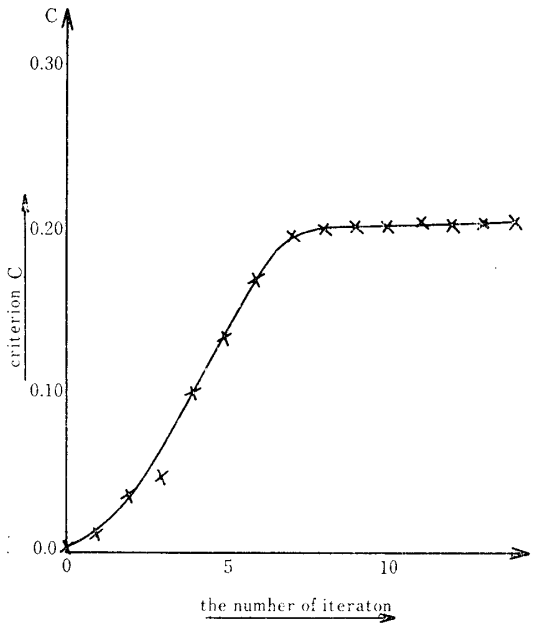


Fig. 2 Variation of criterion C against the number of iteration

Table 1 Inpnt patterns of Class 1

Dimension / Known Class	1	2	3	4	5	6	7	8	9	Normalzed factor
A	0.0	4 0	3.0	1.0	4.0	4.0	3.0	0.0	1.0	68.0
B	2.0	0.0	1.0	2.0	2.0	3.0	2.0	1.0	1.0	28.0
A	0.0	4.0	2.0	2.0	4.0	4.0	2.0	0.0	2.0	64.0
B	1.0	1.0	0.0	1.0	3.0	1.0	4.0	1.0	3.0	39.0
B	3.0	2.0	1.0	4.0	2.0	1.0	3.0	1.0	2.0	49.0
A	0.0	6.0	2.0	2.0	5.0	4.0	3.0	0.0	2.0	98.0
A	0.0	2.0	1.0	0.0	3.0	4.0	3.0	1.0	1.0	41.0
A	0.0	2.0	2.0	1.0	3.0	3.0	2.0	0.0	2.0	35.0
B	2.0	1.0	0.0	3.0	3.0	2.0	2.0	0.0	1.0	32.0
B	3.0	2.0	1.0	3.0	3.0	1.0	2.0	2.0	2.0	45.0
B	0.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	0.0	10.0
A	0.0	1.0	3.0	1.0	3.0	4.0	2.0	0.0	2.0	44.0
A	0.0	4.0	3.0	1.0	4.0	5.0	3.0	0.0	1.0	77.0
B	3.0	1.0	1.0	3.0	2.0	2.0	3.0	0.0	0.0	37.0
B	2.0	0.0	1.0	3.0	3.0	1.0	3.0	2.0	2.0	41.0

Table 2 Input pattens of Class 2

Dimension / Known Class	1	2	3	4	5	6	7	8	9	Normalized factor
A	1.0	2.0	1.0	2 0	3.0	4.0	2.0	1.0	1.0	41.0
B	1.0	2.0	1.0	2.0	1.0	0.0	1.0	0.0	1.0	13.0
A	1.0	2.0	2.0	2.0	4.0	1.0	4.0	1.0	2.0	51.0
B	1.0	2.0	3.0	3.0	4.0	5.0	2.0	0.0	2.0	72.0
A	1.0	2.0	1.0	2.0	3.0	4.0	2.0	1.0	1.0	41.0
B	2.0	1.0	2.0	3.0	2.0	1.0	2.0	1.0	0.0	28.0
B	2.0	1.0	0.0	3.0	3.0	1.0	3.0	1.0	3.0	43.0
A	0.0	3.0	2.0	2.0	3.0	3.0	1.0	0.0	2.0	40.0
A	0.0	2.0	2.0	1.0	3.0	3.0	2.0	0.0	2.0	35.0
A	1.0	3.0	1.0	3.0	4.0	4.0	2.0	1.0	2.0	61.0
B	3.0	2.0	1.0	4.0	2.0	1.0	3.0	1.0	2.0	49.0
B	1.0	1.0	1.0	2.0	2.0	1.0	2.0	2.0	1.0	21.0
B	3.0	1.0	1.0	3.0	2.0	2.0	3.0	0.0	0.0	37.0
A	0.0	2.0	2.0	2.0	2.0	4.0	2.0	0.0	1.0	37.0
A	1.0	2.0	1.0	2.0	3.0	4.0	2.0	1.0	1.0	41.0

Table 3 The result of Calculation, Class 1

Dimension / Known Class	1	2	3	4	5	6	7	8	9	Normalized factor
B	2.0	1.0	0.0	3.0	3.0	1.0	3.0	1.0	3.0	43.0
B	2.0	0.0	1.0	2.0	2.0	3.0	2.0	1.0	1.0	28.0
B	3.0	1.0	1.0	3.0	2.0	2.0	3.0	0.0	0.0	37.0
B	3.0	2.0	1.0	4.0	2.0	1.0	3.0	1.0	2.0	49.0
B	3.0	2.0	1.0	4.0	2.0	1.0	3.0	1.0	2.0	49.0
B	2.0	1.0	2.0	3.0	2.0	1.0	2.0	1.0	0.0	28.0
B	1.0	1.0	1.0	2.0	2.0	1.0	2.0	2.0	1.0	21.0
B	1.0	1.0	0.0	1.0	3.0	1.0	4.0	1.0	3.0	39.0
B	2.0	1.0	0.0	3.0	3.0	2.0	2.0	0.0	1.0	32.0
B	3.0	2.0	1.0	3.0	3.0	1.0	2.0	2.0	2.0	45.0
A	1.0	2.0	2.0	2.0	4.0	1.0	4.0	1.0	2.0	51.0
B	1.0	2.0	3.0	3.0	4.0	5.0	2.0	0.0	2.0	72.0
B	1.0	2.0	1.0	2.0	1.0	0.0	1.0	0.0	1.0	13.0
B	3.0	1.0	1.0	3.0	2.0	2.0	3.0	0.0	0.0	37.0
B	2.0	0.0	1.0	3.0	3.0	1.0	3.0	2.0	2.0	41.0

Table 4 The result of Calculation, Class 2

Dimension / Known Class	1	2	3	4	5	6	7	8	9	Normalized factor
A	1.0	2.0	1.0	2.0	3.0	4.0	2.0	1.0	1.0	41.0
A	0.0	4.0	3.0	1.0	4.0	5.0	3.0	0 0	1.0	77.0
B	0.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	0.0	10.0
A	0.0	1.0	3.0	1.0	3.0	4.0	2.0	0.0	2.0	44.0
A	1.0	2.0	1.0	2.0	3.0	4.0	2.0	1.0	1.0	41.0
A	0.0	6.0	2.0	2.0	5.0	4.0	3.0	0.0	2.0	98.0
A	0.0	4.0	3.0	1.0	4.0	4.0	3.0	0.0	1.0	68.0
A	0.0	3.0	2.0	2.0	3.0	3.0	1.0	0.0	2.0	40.0
A	0.0	2.0	2.0	1.0	3.0	3.0	2.0	0.0	2.0	35.0
A	1.0	3.0	1.0	3.0	4.0	4.0	2.0	1.0	2.0	61.0
A	0.0	2.0	2.0	1.0	3.0	3.0	2.0	0.0	2.0	35.0
A	0.0	2.0	1.0	0.0	3.0	4.0	3.0	0.0	2.0	41.0
A	0.0	4.0	2.0	2.0	4.0	4.0	2.0	0.0	2.0	64.0
A	0.0	2.0	2.0	2.0	2.0	4.0	2.0	0.0	1.0	37.0
A	1.0	2.0	1.0	2.0	3.0	4.0	2.0	1.0	1.0	41.0

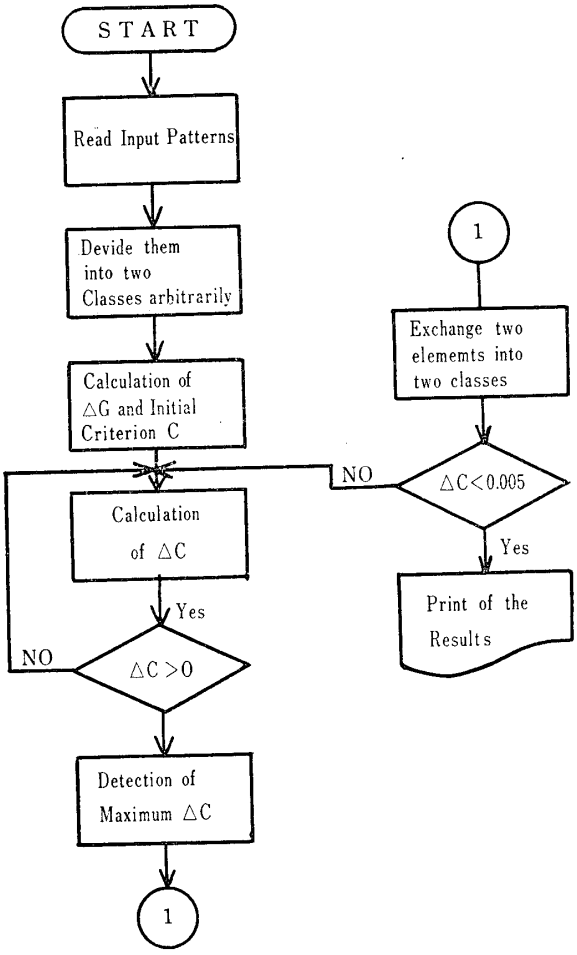


Fig. 3　Flow-Chart

6. 結　　論

以上のことから

1) Karhunen-Loéve 法はパターンの分布が未知である場合にも，教師なしで類別を行なうことができる非常に有力な方法である．

2) 実験結果から(14)式の criterion は十分な機能をはたす．

3) Karhunen-Loéve 法を用いた教師なしの類別法は基本的には距離の概念を導入したものである．

ことが明らかになった．今後は，多くの実験を行ないCの値が最終値に達するまでの変化の様子を検討するつもりである．最後に，日頃から何かと御世話になる工学部平田助教授に深く感謝します．同時に計算機で何かと手助けしていただいた西村女史に感謝します．

参 考 文 献

1) S. Watanabe : Knowing and Guessing PP 380—403 (1969) John Wiley and Sons,

2) K. Fukunaga, et. al : "Application of the Karhune -Loe've Expansion to Feature and Ordering" IEEE Trans. C—19 PP 314—318 (1970, April)

3) 富田，他：Karhunen-Loe've 直交系による教師なしの類別について，インホメーション理論研究資料 IT 70—60 (1970)

4) Y.T. Chien and K.S Fu : "Selection and Ondering of Feature Observation in a Pattern Recognition System" Infomation and Control 12 PP 395—404(1968)

5) 瀬良豊士：相関法を用いたパターン認識　山口大学工学部研究報告，23，1，PP 1—5 (1972)

6) 瀬良豊士：再構成された低次元パターンに関する考察，電気四学会中国支部連合大会，$3_{23}16$ (1972)

(昭和48年4月14日受理)