Unsupervised Clustering in an Orthogonal System

Toyoshi Torioka*

Abstract

It is known that the Karhunen-Lóeve orthogonal system provides a convenient tool to unsupervised clustering problem. Though some results have been obtained along this line, the method is not yet fully applied to the practical pattern clustering problem. The method of unsupervised clustering problem is first explained from the K-L system point of view. Secondly, an algorithm is proposed for the unsupervised clustering problem. Finally, the algorithm is applied to reformed patterns, which are obtained from practical patterns by the method of reformation proposed by the author. Computer-simulated experiments are carried into effect in order to test the usefullness of the present method.

1. Introduction

The present paper gives a method of unsupervised clustering by the use of the K-L system, in order to classfy a set of patterns with an unknown probability distribution.

The principle of the K-L system were used by T.Iijima(1963)[1] in his theory feature extraction. The same concept is made use of from an independent standpoint by S.Watanabe(1965)[2]. Later, the K-L system is applied to the unsupervised clustering problem by K.Fukunaga, et. al.(1970)[3], where the two-categories problem is treated and the equal variance is assumed. A generalised theory of the two-categories problem is given by S.Tomita, et. al.(1971)[4]. In the present paper, it is first shown that the concept of distance plays a foundamental role in the method of unsupervised clustering by the use of K-L system. Secondly, a new algorithm of unsupervised clustering in proposed and is discussed from the basic view point of K-L expansion. Finally, the proposed method is applied to a set of reformed patterns and the results of computer simulated experiments are shown.

This paper first appeared in Japanese in Joho－Shori (Journal of the Information Processing Society of Japan), Vol. 14, No. 10 (1973), pp. 746～753.

* Technical College, Yamaguchi University, Japan.

2. Karhunen-Lóeve Expansion

It is well known that the Karhunen-Lóeve expansion is the most poweful method for extracting the features of patterns. Here the idea of the K-L expansion is introduced in order to explain the fact that the method applied to unsupervised clustering makes use of the distance between two categories. The K-L expansion is a method for expanding a random pattern vector in terms of a set of orthonormal vectors. Now, let $X_i^{(k)}$ be a normalised pattern vector sampled from unknown probability distributions. Then the set of normalized pattern vectors $X_i^{(k)}$'s are denoted by

$S=\{ X_i^{(k)} \mid k=1\sim K \ , \ i=1\sim L \}$ -------------------------(1),

where $\{ a_{il}^{(k)} \}$ are the coefficients of expansion and a m is the dimension of the pattern $X_i^{(k)}$. K is the number of pattern classes and L is the number of patterns By a system $\{ \xi_l \}$ of orthonormal vectors, a pattern vector $X_i^{(k)}$ is expanded as follows,

$X_i^{(k)} = \sum_{l=1}^{m} a_{il}^{(k)} \cdot \xi_l \qquad a_{il}^{(k)} = (X_i^{(k)} \cdot \xi_l)$ -------------------------(2).

Let p be the probability of occurrence of a pattern in class k. Then the autocorrelation matrix G of the patterns of all the classes is calculated by

$G=\sum_{k=1}^{K} p_k \cdot E(X_i^{(k)} \cdot X_i^{(k)T})$ -------------------------(3),

where $E(X_i^{(k)} \cdot X_i^{(k)T})$ is the autocorrelation matrix of class k. Next, let

$\{ \lambda_l \mid l=1\sim m \}$ -------------------------(4),

be the set of eigenvalues $\{ \lambda_l \}$ of G. The eigenvectors of G, i.e, the solutions of Eq.(5)

$G \cdot \xi_l = \lambda_l \xi_l$ -------------------------(5),

forms a set of orthonomal vectors $\{ \xi_l \}$ and is called the Karhunen-Lóeve system. Moreover, eigenvalue $\{ \lambda_l \}$ is represented by

$\lambda_l = \xi_l^T G \xi_l = \sum_{k=1}^{K} p_k \cdot \mathrm{Var}(a_{il}^{(k)})$ -------------------------(6),

where $\mathrm{Var}(a_{il}^{(k)})$ is the variance of $a_{il}^{(k)}$. Hence, the eigenvalue λ_l is considered as the degree of importance of the vector ξ_l , in measuring featres of patterns.

3. Criterion of Unsupervised Clustering

An explanation is given to a problem of classifing a set of patterns sampled from two unknown probability distributions into two classes. The set S is denoted, in this case, by

$S=\{ X_i^{(k)} \mid k=1,2 \ , \ i=1\sim L \}$ -------------------------(7).

Now, it become necessary to define a criterion by which at first, patterns in the set S are classified into two classes without supervision. Let us classify the patterns arbitrarily and let M and N be the numbers of patterns in the two classes. Patterns

$X_i^{(1)}$ and $X_i^{(2)}$ are expanded as follows

$$X_i^{(1)}=\sum_{\ell=1}^{m} a_{i\ell}^{(1)}\,\xi_\ell \qquad X_i^{(2)}=\sum_{\ell=1}^{m} a_{i\ell}^{(2)}\,\xi_\ell \qquad \text{------------------------(8).}$$

The autocorrelation matrices G_1 and G_2 of the both classes are expressed by

$$G_1=\frac{1}{M}\sum_{i=1}^{M} X_i^{(1)}\cdot X_i^{(1)T} \qquad G_2=\frac{1}{N}\sum_{i=1}^{N} X_i^{(2)}\cdot X_i^{(2)T} \qquad \text{------------------------(9),}$$

and their eigenvalues are denoted by

$$\left\{\lambda_\ell^{(1)} \,\middle|\, \ell=1\sim m\ ,\ \lambda_\ell^{(1)}=\frac{1}{M}\sum_{i=1}^{M}(a_{i\ell}^{(1)})^2\right\} \qquad \left\{\lambda_\ell^{(2)} \,\middle|\, \ell=1\sim m\ ,\ \lambda_\ell^{(2)}=\frac{1}{N}\sum_{\ell=1}^{N}(a_{i\ell}^{(2)})^2\right\} \ \text{---(10),}$$

respectively. The criterion of unsupervised clustering is to maximize C defined by

$$C=\sum_{\ell=1}^{m}(\lambda_\ell^{(1)}-\lambda_\ell^{(2)})^2 \qquad \text{------------------------(11),}$$

where C is the squared sum of the differences between corresponding feature measurements of two classes and it may be regarded as the distance between two classes. From this definition, it is concluded that the better the result of clustering becomes, the larger the value of C becomes. The optimum classification is obtained when the value C becomes the largest. It is, however, very difficult for us to calculate $\{\lambda_\ell^{(1)}\}$ and $\{\lambda_\ell^{(2)}\}$ directly from Eq.(10) and then calculate C from (11). In order to simplify the procedure of calculating C, let G_0 be the matrix obtained by substrating G_2 from G_1. Then, the trace of G_0 is calculated as follows

$$\mathrm{tr}(G_0)=\mathrm{tr}(G_1-G_2)=\mathrm{tr}\begin{bmatrix}\lambda_1^{(1)}-\lambda_1^{(2)} & & 0\\ & \ddots & \\ 0 & & \lambda_m^{(1)}-\lambda_m^{(2)}\end{bmatrix} \qquad \text{------------------------(12).}$$

From Eq.(12), the criterion C is simply calculated by

$$C=\sum_{\ell=1}^{m}(\lambda_\ell^{(1)}-\lambda_\ell^{(2)})^2=\sum_{\ell=1}^{m}(\lambda_\ell^{(0)})^2=\mathrm{tr}(G_0)^2 \qquad \lambda_\ell^{(0)}=\lambda_\ell^{(1)}-\lambda_\ell^{(2)} \qquad \text{--------------------(13).}$$

Therefore, the optimum classification is obtained by the method of making the trace of G_0 maximum without calculating the $\{\lambda_\ell^{(1)}\}$ and $\{\lambda_\ell^{(2)}\}$.

4. A Optimization Method

The criterion value C of unsupervised clustering has been given by Eq.(13). It is related to the feature measurements of input patterns obtained by K-L expansion. Practically, it is needed to exchange sequentially patterns between two classes, in order to increase the value C until the optimal value is obtained.

An algorithm, shich exchanges two patterns of the two classes has been proposed in reference [2], but this algorithm has two shortcomings as follows. One of them is the restriction that the number of patterns should be the same in each class. The other is a fact that one misclassification brings inevitably about two error classifications. A improved algorithm is mentioned below. It consists of two subalgorithms, one exchanging two patterns of the two classes at the same time and the other adding a pattern to one class from the other class. These subalgorithms are hereafter called the exchange algorithm and the addition algorithm, respectively.

Now, let the set of patterns in Eq.(7) be classified arbitrarily into two calsses, M and N being the pattern numbers of the both classes. The matrix G_0 is calculated by Eq.(9) and the value of C is obtained from Eq.(13). Next, a mathematical expression is given to the exchange and addition algorithms and the change of C acompanying the algorithms is obtained. Let $X_r^{(1)}$ and $X_t^{(2)}$ be patterns exchanged fistly. By the exchange of the two patterns, the autocorrelation matrices G_1 and G_2 change to

$$G_1=\frac{1}{M}\sum_{i=1}^{M}X_i^{(1)}\cdot X_i^{(1)T}-\left(\frac{1}{M}X_r^{(1)}\cdot X_r^{(1)T}-\frac{1}{M}X_t^{(2)}\cdot X_t^{(2)T}\right) \text{------------------------}(\ 14\ ),$$

$$G_2=\frac{1}{N}\sum_{i=1}^{N}X_i^{(2)}\cdot X_i^{(2)T}-\left(\frac{1}{N}X_t^{(2)}\cdot X_t^{(2)T}-\frac{1}{N}X_r^{(1)}\cdot X_r^{(1)T}\right) \text{------------------------}(\ 15\ ),$$

and the new matrix $G_0'=G_1-G_2$ is denoted as

$$G_0'=G_0+\left(\frac{M+N}{M\cdot N}\right)\cdot\Delta G \qquad \Delta G_0=\left(X_t^{(2)}\cdot X_t^{(2)T}-X_r^{(1)}\cdot X_r^{(1)T}\right) \text{-----------------}(\ 16\ ).$$

Hence, the increment ΔC of the value C is given by

$$\Delta C=tr(G_0')^2=2\left(\frac{M+N}{M\cdot N}\right)tr(G_0\cdot\Delta G)+\left(\frac{M+N}{M\cdot N}\right)tr(\Delta G_0)^2 \text{----------------------}(\ 17\ ).$$

Accordingly, in the case of $\Delta C>0$, the classification of the patterns becomes better than earlier, if $X_r^{(1)}$ and $X_t^{(2)}$ are exchanged. In case $\Delta C<0$, they should not be exchanged. Let us consider the case in which a pattern $X_r^{(1)}$ of class 1 is added to class 2. By the addition of pattern $X_r^{(1)}$, the matrices G_1 and G_2 are rewritten as

$$G_1=\frac{M}{M-1}\left(\frac{1}{M}\sum_{i=1}^{M}X_i^{(1)}\cdot X_i^{(1)T}-\frac{1}{M}X_r^{(1)}\cdot X_r^{(1)T}\right) \text{-------------------------}(\ 18\ ),$$

$$G_2=\frac{N}{N+1}\left(\frac{1}{N}\sum_{i=1}^{N}X_i^{(2)}\cdot X_i^{(2)T}+\frac{1}{N}X_r^{(1)}\cdot X_r^{(1)T}\right) \text{-------------------------}(\ 19\ ),$$

and the matrix G_0' is calculated as

$$G_0=G_1-G_2=\frac{M\cdot N}{(M-1)(N+1)}G_0+\Delta G_0$$

$$\Delta G_0=\frac{1}{(M-1)(N+1)}\left(\sum_{i=1}^{M}X_i^{(1)}\cdot X_i^{(1)T}+\sum_{i=1}^{N}X_i^{(2)}\cdot X_i^{(2)T}-(M+N)X_r^{(1)}\cdot X_r^{(1)T}\right) \text{------------}(\ 20\ ).$$

Hence, the increment ΔC is given by

$$\Delta C=tr(G_0')^2-tr(G_0)^2=2tr(G_0\cdot\Delta G_0)+tr(\Delta G_0)^2 \text{------------------------}(\ 21\ ),$$

which is similar to Eq.(17), provided M and N are large. Conversely, in case of the addition of pattern $X_t^{(2)}$, the matrix G_0' is calculated as

$$G_0'=\frac{1}{(M+1)(N-1)}\cdot G_0+\Delta G_0$$

$$\Delta G_0=\frac{1}{(M+1)(N-1)}\left\{(M+N)X_t^{(2)}\cdot X_t^{(2)T}-\left(\sum_{i=1}^{M}X_i^{(1)}\cdot X_i^{(1)T}+\sum_{i=1}^{N}X_i^{(2)}\cdot X_i^{(2)T}\right)\right. \text{---------}(\ 22\ ),$$

and the increment ΔC is similarly given. Therefore, if the increment ΔC becomes positive, pattern $X_r^{(1)}$ (or $X_t^{(2)}$) is added to class 2 (or class 1) on the analogous reason at the exchange algorithm. Moreover, the second terms in Eqs.(17) and (21) can be neglected and ΔC can be approximated by

$$\Delta C=2\left(\frac{M+N}{M\ N}\right)tr(G_0\cdot\Delta G) \qquad \Delta C=2tr(G_0\cdot\Delta G) \text{-------------------}(\ 25\ ),$$

because these terms are always positive. The complete algorithm is easily obtained from the explanation described above, but is omitted here. Refer to [6] for the detail of the algorithm.

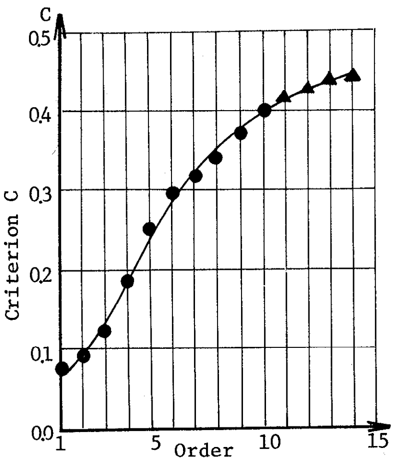

Fig.1 Variation of Criterion, C

5. Results of Experiments

In computer-simulated experiments, two sets of patterns were prepared. One of them consists of about 30 reformed patterns from the two letters A and B, and the other those from E and F. Two experiments are preformed under the following two conditions:

Case(1) Patterns of the two classes have high statistical dependency.

Case(2) Patterns of the two classes have low statistical dependency.6)

The results of the experiments are shown in Table 1 and are fairly satisfactory. It is also turned out that the large the value of the degree of dependence becomes, the better the result of classification becomes. In Fig. 1, the value of criterion C is plotted in its increasing order.

Table 1 Results of the Classification

Initial State					Final State				Misclass-fication
Name Case	Set of Patterns	Patten	Class1	Class2	Set of Patterns	Pattern	Class1	Class2	
(1)	A B	A	10	6	A B	A	16	0	0
		B	8	9		B	2	15	2
	E F	E	10	6	E F	E	12	4	4
		F	8	9		F	5	12	5
(2)	A B	A	7	11	A B	A	18	0	0
		B	10	5		B	1	14	1
	E F	E	10	8	E F	E	18	0	0
		F	6	11		F	0	17	0

The marks of circles and triangles in the figure indicate that the exchange algorithm and addition algorithm, respectively, are employed in order to increase ΔC.

6. Conclusion

The following four conclusions are obtained from the above results

1) The method of unsupervised clustering by the use of K-L system consists of maximizing the criterion C, which is nothing but the Euclidean distance of two sets of patterns defined by the feature measurements.

2) The criterion is simplified to Eq.(13) by means of the concept of the trace,

and the trace is calculated from the square of the matrix G_σ.

3) The present method is a poweful method for unsupervised clustering, even if the probability ditribution is unknown and different.

4) Some shortcomings of the exchange algorithm proposed by K.Fukunaga, et. al. being overcome, the present method is improved to given a better classification by the aid of the addition algorithm.

7. Acknowlegement

The author wishes to thank Prof.S.Amari of the University of Tokyo and Prof. T.Hirata of the Yamaguchi University for their kind discussion and valuable comments. He also thanks Miss Y.Nishimura for helping him in computer programming.

References

1. T.Iijima, "Theory of pattern recognition," Jour. Inst. Electronics. Comm. Engrs. Japan, Vol.46, No.1, pp.1582-1590, 1963.
2. S.Watanabe, "Knowing and Guessing," pp.380-403, Wiley, 1969.
3. K.Fukunaga and W.L.G.Koontz, "Application of the Karhunen-Lóeve Expansion to Feature Selection and Ordering," IEEE Trans. C-19, pp.311-318, 1970.
4. S.Tomita,et.al., "Classification for Patterns by the Karhunen-Lóeve Orthogonal System Without Supervision," Trans. Inst. Electronics Comm. Engrs. Japan, Vol.54-C, No.8, pp.767-774, 1971.
5. T.Sera, "Pattern Recognition by the Codes of 3x3 Elements with Karhunen-Lóeve Orthogonal System," Infor. Proc. Japan, Vol.13, No.4, pp.210-217, 1972.
6. T.Sera, "Unsupervised Clustering in an Orghogonal System," Infor. Proc. Japan, Vol.14, No.10, pp.746-753, 1974.