A NEW CLUSTERING APPROACH FOR PATTERN RECOGNITION IN UNSUPERVISED ENVIRONMENT

BELUR V. DASARATHY
(*School of Automation, Indian Institute of Science, Bangalore 560012, India*)

Received on October 10, 1973 and in revised form on February 25 1974

ABSTRACT

A practical clustering technique for learning in an unsupervised mode, based on a new concept of measure of closeness of samples, has been developed. This new approach is found to be capable of zeroing on the truly inherent clusters of the different unspecified pattern classes with a very high degree of certainty.

Key words: Pattern recognition; Clusters in patterns.

Basic to the concept of clustering techniques [1–4], in the field of unsupervised learning for pattern recognition, is the need for defining a set of initial cluster centers. These cluster centers, around which the clusters are formulated, are to be representative of the different clusters or pattern classes arising in the pattern recognition problem. One among the numerous methods of arriving at such clustering initiation points is that of determining a pre-specified (depending on the number of clusters desired or expected) number of samples which are considered to be "farthest from one another [1] " in the Euclidean sense. The tacit assumption underlying such an approach is that the samples, that are farthest from one another, necessarily belong to different clusters. Otherwise the ensuing clustering process may not lead to the true inherent pattern clusters. But such an assumption, while convenient, is certainly questionable and one can easily visualize environments wherein such assumptions may not be valid. As can be gleaned from Fig. 1, the Euclidean distance between the samples, $X^{(1)}$ and $X^{(2)}$, belonging to the same cluster C_1, is greater than the distance between the samples, $X^{(1)}$ and $X^{(3)}$ or $X^{(2)}$ and $X^{(3)}$, belonging to the two different clusters C_1 and C_2. Hence, the method of arriving at representative cluster samples on the basis of maximum intersample Euclidean distance measure being the direct vectorial sum of the distances along each feature direction, does not take into account the differences in the spread of the data along the different feature directions. Thus an abnormally high spread in the data along one direction even *within* one cluster can completely overwhelm the

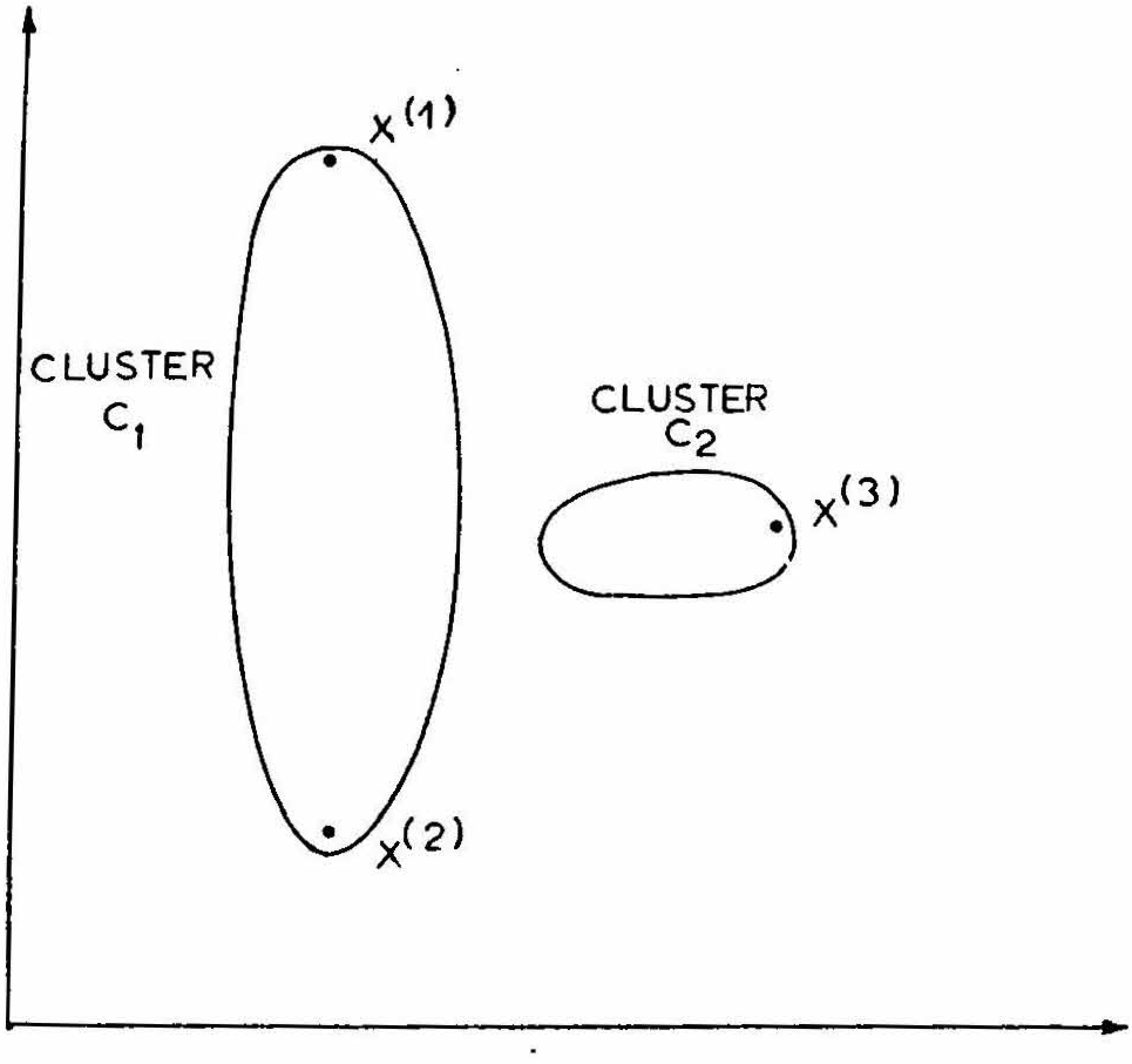

FIG. 1

effect of fairly significant spreads possible *between* the two clusters along the other feature directions. A clustering process initiated by such blatantly false representative points of the two clusters cannot be expected to lead to truly inherent and stable clusters.

To obviate this deficiency in the Euclidean distance measure, a new normalized measure of closeness of samples is proposed. If instead of using the absolute Euclidean distance between the samples as the measure of closeness, one were to determine the distance between the samples along each direction, normalized with respect to the maximum spread in the data along that direction and vectorially sum up these normalized distances along the different feature directions, a truer measure of " farthest from one another " can be expected. In general, such a measure can, with a higher degree of expectancy, lead to the samples representative of the different clusters. The samples so arrived at form the set of clustering initiation points. The rest of the sample vectors are assigned to the different clusters so defined on the basis of closeness to these clustering initiation points. As before, this measure of closeness, instead of being based on the absolute Euclidean distances, is defined in terms of the vectorial sum of the normalized distances along each feature direction, the normalization being relative to the maximum spread of the data along the corresponding feature directions. After the first iteration of this clustering process, a new set of cluster centers are determined as the center points of all the different sample vectors assigned to the corresponding clusters. However, the reassignment procedure of the sample on the basis of closeness to these new cluster centers is now modified to take into account maximum information that can be derived from the clusters as developed at this stage. The normalization factors, instead of being the maximum spread (in each feature direction) over the complete data set, is defined to be the maximum spread (in each feature direction) over the samples belonging only to the individual cluster to whose center the closeness measure is being evaluated. This clustering process of determining the cluster centers and assigning, the sample vectors on the basis of this modified measure of closeness is carried on recursively till truly inherent and stable clusters are obtained. This procedure is described in detail in the sequel.

Let $X = \{X^j = \{x_i^j : i = 1, 2 \ldots n\} : j = 1, 2 \ldots p\}$ be the set of sample vectors which is to be partitioned into m cluster sets $(m \ll p)$.

Now,

D^{jk} : Measure of closeness between sample vectors X^j and X^k

$$\underline{\underline{\triangle}} \sum_{i=1}^{n} [(x_i^j - x_i^k)/\triangle x_i]^2 \tag{1}$$

where,

$$\begin{aligned} \triangle x_i &= (x_i^{\max} - x_i^{\min}) \\ x_i^{\max} &= \max_{X^j \epsilon X} [x_i^j] \\ x_i^{\min} &= \min_{X^j \epsilon X} [x_i^j] \end{aligned} \tag{2}$$

The samples X^{J_1} and X^{J_2} are said to be farthest apart and representative of the two clusters that are farthest apart, where J_1 and J_2 are given by

$$(J_1, J_2) = (j, k) \ni \begin{bmatrix} D^{jk} \\ (j, k = 1, 2 \ldots p) \end{bmatrix} \text{ is maximum} \tag{3}$$

Further sample vectors $\{X^{J_3}, \ldots X^{J_k}, \ldots X^{J_m}\}$ representative of the other $(m - 2)$ pattern clusters are given by X^{J_k}:

where,

$$J_k = j \ni \left[\begin{array}{l} \left(\sum_{r=1}^{k-1} D^{jJ_r}\right) \\ j = (1, 2 \ldots p) \\ \quad \neq (J_1, J_2, \ldots J_{k-1}) \end{array}\right] \text{ is maximum} \tag{4}$$

$(3 \leq k \leq m)$

Equation (4) is repetitively applied till all the m requisite cluster initiation points given by the sample vector set $\{X^{J_1}, \ldots X^{J_m}\}$ are determined.

Now a sample vector X^j is assigned to the cluster C_r,

where

$$r = k \ni \left[\begin{array}{l} D^{jJ_k} \\ (k = 1, 2 \ldots m) \end{array}\right] \text{ is minimum} \tag{5}$$

The new cluster centers $\bar{X}^k$ $(k = 1, 2, \ldots m)$ are determined as the centers of all samples assigned to the cluster C_k and is given by $\bar{x}_i{}^k : (i = 1, 2 \ldots n)$ where

$$\bar{x}_i{}^k = \left[\sum_{X^j \epsilon C_k} x_i^j / p_k\right] \tag{6}$$

with p_k being the number of sample vectors assigned to the cluster C_k.

The process of reassignment of the samples is now carried out on the basis of a modified measure of closeness. This new measure takes full advantage of the information available at the end of the initial stage of this recursive learning scheme by determining the actual maximum spread of each cluster (along each feature direction) separately and using it as the basis for normalizing the Euclidean distance measured relative to the corresponding cluster centre, *i.e.*, $\mathscr{D}^{jk}$: Modified Measure of closeness

$$\triangleq \sum_{i=1}^{n} [(x_i^j - \bar{x}_i{}^k)/\Delta^{(k)} x_i]^2 \tag{7}$$

$$\begin{aligned} \Delta^{(k)} x_i &= (x_i{}^{\max} - x_i{}^{\min})|_k \\ x_i{}^{\max}|_k &= \max_{X^j \epsilon C_k} [x_i^j] \\ x_i{}^{\min}|_k &= \min_{X^j \epsilon C_k} [x_i^j] \end{aligned} \tag{8}$$

and a sample vector X^j is assigned to the cluster C_r where

$$r = k \ni \left[\begin{array}{l} \mathscr{D}^{ik} \\ (k = 1, 2, \ldots m) \end{array}\right] \text{ is minimum}$$

This clustering process eqn (6) to eqn (9) is carried out recursively till stable clusters $\{C_k : k = 1, 2, \ldots m\}$ are obtained, *i.e.*, the cluster centers remain

TABLE I

Results of the new clustering process as applied to Iris data

	Cluster initiation points chosen by the normalized measure of closeness				FOR CLUSTER C_1				FOR CLUSTER C_2			
					x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
					5·00	2·00	3·50	1·00	7·70	3·80	6·70	2·20
Iteration Number	Number of samples assigned to the cluster				CENTER OF CLUSTER C_1				CENTER OF CLUSTER C_2			
	C_1		C_2		x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
1	54	(42)	46	(38)	5·850	2·681	4·352	1·387	6·746	3·096	5·537	2·015
2	50	(44)	50	(44)	5·850	2·700	4·284	1·334	6·674	3·044	5·528	2·018
3	50	(46)	50	(46)	5·872	2·722	4·276	1·320	6·652	3·022	5·536	2·032
4	50	(46)	50	(46)	5·886	2·720	4·276	1·316	6·638	3·024	5·536	2·036
	Centers of the pattern classes as computed from the given two class Iris data using the information (of their class labels) withheld to the unsupervised learning scheme				5·936	2·770	4·260	1·326	6·588	2·974	5·552	2·026

Figures in parentheses refer to the actual number of these samples known to belong to the corresponding pattern class. This knowledge is derived from the class label information available for the samples but withheld to the clustering process in order to simulate the unsupervisd learning environment.

unchanged in successive iterative stages. The convergence properties of such recursive procedures have been studied in the literature [6].

This modification to the normalized measure of closeness is not applicable at the start of the clustering process as no clusters, however crudely formed, are available. However, this modified measure makes a significant contribution to the clustering process in that, unlike other existing• clustering techniques, it derives maximum beneficial information out of the clusters as developed at the previous stage of clustering process.

With a view to assess the efficacy of this new clustering technique, the relevant computational scheme was implemented on the IBM 360/44 and tested against the well-known Iris data [5]. This two-class data set consisting of one hundred (four-dimensional) sample vectors was fed in withholding the available information regarding the class labels of the individual samples to simulate the unsupervised learning environment. Table I brings out the performance of the resulting iterative clustering process. Listed at the top of the table are the two cluster initiation points determined as the farthest-apart-samples in the sense of the normalized measure of closeness. The results of the successive iterations, listed thereunder, reveal the progressive stabilization occurring in the cluster formation process leading to the inherently stable clusters at the end of four iterations. The iteration scheme comes to a halt when no more changes are noted in the position of the center of each of these two clusters.

That, in this example, the number of samples assigned to each cluster turned out to be equal, is admittedly per chance. Still, the results convincingly bring out the significance and merit of this new measure of closeness. This is particularly clear from the fact that among the samples assigned to each cluster, the actual number of samples that truly belong to the corresponding pattern class increases as the iterations proceed and level off at a significantly high percentage (92%) at the end of the iteration process. This information although not made use of in the clustering process to simulate the unsupervised mode of learning, comes in handy in the *a-posteriori* evaluation of this new technique.

REFERENCES

[1] Nagy, G. .. The application of non-supervised learning to character recognition in *Pattern Recognition*, Ed. Kanal, L., Thompson Book Co., 1968, pp. 391–398.

[2] Haralick, R. M. and Kelly, G. L. Pattern recognition with Measurement space and spatial clustering for multiple images. *Proc. IEEE.*, 1969, 57 (4), 654–665.

[3] Haralick, R. M. and Darling, G. *Non-parametric Unsupervised Learning: Ideas and Results.* Second Hawaii International Conference or System Science, University of Hawaii, 1968.

[4] Haralick, R. M. .. Adaptive pattern recognition of agriculture in Western Kansas by using a predictive model in construction of similarity sets. *Proceedings of the 5th Symposium on Remote Sensing of Environment, University of Michigan*, 1967, pp. 343–355.

[5] Fischer, R. A. .. The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, 1936, 7, 179–186.

[6] MacQueen, J. .. Some methods for classification and analysis of multivariate observations. *Proceedings of Berkeley Symposium on Probability and Statistics*, 1967, pp. 281-297.