INT. J. SYSTEMS SCI., 1978, VOL. 9, NO. 8, 887–897

Adaptive learning algorithm in classification of fuzzy patterns An application to vowels in CNC context

S. K. PAL†, A. K. DATTA† and D. DUTTA MAJUMDER†

An adaptive algorithm for recognition of ill-defined patterns using weak representative points and single pattern training procedure is presented from the standpoint of fuzzy set theory. The method includes both supervised and non-supervised schemes. A non-adaptive algorithm with fixed reference and weight vectors is also described to describe the efficiency of the system's adaptiveness to a new input.

This was implemented to machine recognition of vowel sounds of a number of speakers in Consonant–Vowel Nucleus–Consonant (CNC) context considering the first three vowel formants as input features. The decision of the machine is governed by the maximum value of fuzzy membership function. A recognition rate, particularly for weak initial representative vectors, was seen to be dependent on the sequence of incoming patterns. As the process of classification continued, the learned mean vectors approached their respective true values of the clusters. Again, once the optimum size of training set is obtained, the role of the external supervisor became insignificant.

1. Introduction

Learning is a process which improves the system's performance by acquiring necessary information for decision during the system's operation. The decision in the process is based on the information learned (estimated) and obtained from the observed patterns and if the information learned gradually approaches the true information, then the decision will eventually approach the optimal decision as if all the desired information of each pattern class is known. Therefore, the performance of the system in classifying a pattern during the system's operation is gradually improved. Depending upon whether the correct classification of the input patterns observed are known or not, the learning process performed by the system can be classified into 'supervised learning' and 'non-supervised learning'. In non-supervised learning the correct classification of the observed patterns is not available, i.e. the patterns used are unlabelled and the estimation of the unknown parameters must be performed according to its own decision instead of the labels given by a teacher. In other non-supervised machines the problem of learning is often reduced to a process of successive estimation of some unknown parameters in a mixture distribution of all possible pattern classes (Patric and Hancock 1966, Chien and Fu 1967). For supervised learning, Bayesian estimation (Abramson and Braverman 1962) and stochastic approximation (Fu *et al.* 1966) can be used to successively learn unknown parameters in a given form of feature distribution of each class. Training procedure based on least mean square error (Koford and Groner 1966) and error correcting technique (Duda and Fossum 1966) has also been proposed for linear classifiers and piecewise linear classifiers respectively. Ho and Kasyap (1965) and

Received 1 September 1977.

† Electronics and Communication Sciences Unit, Indian Statistical Institute, 203 B T Road, Calcutta 700 035, India.

Wee and Fu (1968) suggested a few learning methods where the input training patterns, instead of applying sequentially, were used in groups. Group pattern training procedures as compared to single pattern training procedures do involve the increase of computations and storage requirements but converge to the optimum reference vectors in lesser number of iterations (Fu 1968).

The problem of speech recognition involves multilevel decision processes (Reddy 1975, 1976, Lindgren 1965, 1967, Klatt and Stevens 1973, Weinstein *et al.* 1975, Dutta Majumder and Datta 1969, Sharma and Yegnanarayana 1977, Trans. I.E.E.E. 1975, Proc. I.E.E.E. 1976) ranging from recognition of vowel, consonant, isolated word by a single speaker and limited vocabulary system for a few trained speakers to a connected speech recognition system with unlimited vocabulary for a large number of speakers and speech-understanding systems (Woods and Makhoul 1974, Reddy 1976). Since speech is a pattern of biological origin and carries information regarding the message, the speaker, his health and mood, it is found to a considerable extent to be fuzzy in nature. There exists no precise boundary, due to inherent vagueness (fuzziness) rather than randomness in the patterns. Again, since the conditional densities of classes are not known and only a small number of design samples are available, the classifiers based on similarity or dissimilarity measure within the framework of fuzzy language theory (Zadeh *et al.* 1975) appear to be suitable to their recognition (Dutta Majumder and Pal 1977 a, Pal and Datta Majumder 1977 a, b). Of course, both stochastic and fuzzy techniques of classification can be derived from two different constraints in probabilistic concepts, namely : statistical independence (stochastic) and logical implication (fuzzy) (Gaines 1975, Stallings 1977). Zadeh (1968) proposed a probability measure over fuzzy events where the probability of a fuzzy event is equal to the expected value of its membership function.

The present paper confines itself to demonstrating the adaptive efficiency of a system in non-supervised and supervised recognition methods of Indian Telugu vowel sounds in CNC combination starting with the weakest possible representative points in each of the multidimensional classes. The machine was initially trained only with five utterances of one of the three speakers and unknown patterns were inserted sequentially into the system for classification. The test set does contain about 900 such utterances, constituting a three-dimensional vector space. Each of the dimensions is represented by one of the first three vowel formant frequencies extracted from spectrographic analysis. Similarity measures in a classifier are based on the fuzzy membership value. The second part of the experiment consists of 20 initial learning samples selected randomly from each of the vowel categories to see the effect of the increased number of training patterns on the system performance. A general purpose digital computer, Honeywell 400, was used for analysis.

2. Decision rule and learning algorithm

Consider an N-dimensional feature vector space Ω_X containing m ill-defined pattern classes to be recognized with a defined set of N-dimensional prototypes $R_1, R_2, \ldots, R_j, \ldots, R_m$, such that

$$R_j^{(l)} \in R_j$$

$l = 1, 2, \ldots, h_j$, h_j is the number of reference vectors in set R_j.

2.1. *Fuzzy membership function*

The decision of the classifier for the purpose of recognition of an unknown pattern $X=[x_1, x_2, \ldots, x_n, \ldots, x_N]$ is based on the magnitude of its fuzzy membership function corresponding to the jth $(j=1, 2, \ldots, m)$ class :

$$\mu_j(X)=\left(1+\left(\frac{d(X, R_j)}{F_d}\right)^{F_e}\right)^{-1} \tag{1}$$

where F_e is the ' exponential fuzzifier ', F_d is the ' denominational fuzzifier ' and

$$d(X, R_j)=\min_l \|X-R_j{}^{(l)}\|$$

with

$$\|X-R_j{}^{(l)}\|=\left(\sum_n\left(\frac{x_n-\bar{x}_{jn}{}^{(l)}}{\sigma_{jn}{}^{(l)}}\right)^2\right)^{0\cdot5}, \; n=1, 2, \ldots, N$$

denoting the weighted Euclidean distance (Sebestyen 1962, Meisel 1972) between the unknown pattern X and the lth reference vector $R_j{}^{(l)}$ in the jth class in which $\bar{x}_{jn}{}^{(l)}$ and $\sigma_{jn}{}^{(l)}$ correspond to lth prototype and denote the mean and standard deviation of the features along the nth coordinate in the jth class. The fuzzifiers have the effect of altering the ambiguity in a set and hence the overall recognition score (Pal and Dutta Majumder 1977 a, b, Dutta Majumder and Pal 1977 a).

The membership function is defined in such a way that it maps the N-dimensional feature space into an m-dimensional membership space which is a unit hypercube and should satisfy the following conditions :

(i) $\mu_j(X)\to 0$ as $d(X, R_j)\to\infty$

(ii) $\to 1$ $\to 0$

and

(iii) increases decreases

Therefore the membership function $\mu_j(X)$, having a positive value in the interval [0, 1], denotes the degree to which an event X may be a member of or belong to the jth class and the classificatory decision rule would be as follows :

$$\text{decide}: \; X \in C_k \quad \text{if} \quad \mu_k(X)>\mu_j(X), \quad j=k=1, 2, \ldots, m, \quad j\neq k$$

2.2. *Iterative algorithm for parameter estimation*

The component of the reference vector and the weighted vector for a class used in the decisional algorithm are respectively the means and reciprocal of standard deviations of the components of the feature vectors. The reciprocal of the standard deviation is found to provide appropriate phase weights to patterns for their proper classification (Dutta Majumder *et al.* 1976, Pal and Dutta Majumder 1977 a, b).

The basic idea of the recognition system is to draw unknown samples randomly one after another and build up their appropriate classes. The samples that are inserted to a given class modify the centres and relative weights on the axes of the classes. The iterative procedure adopted here is therefore the ' Centre-Variance adjustment algorithm ' in which the weighted

Euclidean distance used in the membership function reflects the ellipsoidal shape of each cluster.

If $\bar{x}_{n(t)}$ and $\sigma_{n(t)}{}^2$ represent the mean and variance of a class along the nth coordinate axis, estimated by first t samples, we note

$$\bar{x}_{n(t)}=\frac{1}{t}\sum_{i=1}^{t} x_i \tag{2 a}$$

and

$$\begin{aligned}\sigma_{n(t)}{}^2&=\frac{1}{t}\sum_{i=1}^{t} (x_i-\bar{x}_{n(t)})^2\\ &=\frac{1}{t}\sum_{i=1}^{t} x_i{}^2-\bar{x}_{n(t)}{}^2\\ &=\frac{1}{t}\,C_{n(t)}-\bar{x}_{n(t)}{}^2\end{aligned} \tag{2 b}$$

where

$$C_{n(t)}=\sum_{i=1}^{t} x_i{}^2$$

Let another sample $x_{(t+1)}$ fall into this class. Then the mean and variance are adjusted as follows :

$$\bar{x}_{n(t+1)}=\frac{t}{t+1}\,\bar{x}_{n(t)}+\frac{1}{t+1}\,x_{(t+1)} \tag{3 a}$$

$$C_{n(t+1)}=C_{n(t)}+x_{(t+1)}{}^2 \tag{3 b}$$

and

$$\sigma_{n(t+1)}{}^2=\frac{1}{t+1}\,C_{n(t+1)}-\bar{x}_{n(t+1)}{}^2 \tag{3 c}$$

Equations (3) provide us with an iterative algorithm for automatic estimation of the mean and variance vectors, given successive samples.

3. Recognition procedure and experimental results

A vocabulary consisting of 600 Telugu words were uttered by three speakers and recorded on an AKAI 1710 tape recorder. The spectrographic analyses were done on a Kay sonagraph model 7029 A. Formant frequencies F_1, F_2 and F_3 for the ten vowels (∂, a :, i, i :, u, u :, e, e :, o and o :) were obtained manually at the steady state of the vowels. Whenever, due to the extreme shortness of the vowels, steady state was not observed, the measurements were taken at the point of congruence of the off-glide and on-glide. The number of samples obtained after processing the spectrograms are only 871. The details of the experimental set-up for recording and having a spectrographic display of the Telugu words, including the nature of speakers, context of words, measurement procedure and extraction of formant frequencies, are presented in our earlier papers (Dutta Majumder *et al.* 1976, Pal and Dutta Majumder 1977 a).

The distribution of the samples uttered by the three speakers in the F_1–F_2 plane of the vector space is sketched in Fig. 1, where the boundaries among vowel classes are seen to be ill-defined.

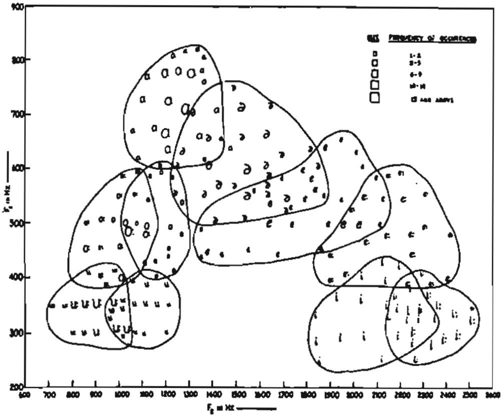


Figure 1. Distribution of Telugu vowels in the F_1–F_2 plane.

A flow-chart for the recognition scheme is shown in Fig. 2. The method consisted mainly of the adaptive recognition of vowels by non-supervised and supervised learning to show the effectiveness of an external teacher in providing supervision and labelling patterns. A non-adaptive scheme with fixed mean and weighted vectors is also presented to demonstrate the efficiency of the system adaptiveness to the new input events.

First, only five samples for each of the classes and uttered by a single speaker were selected from the sample space for initial training of the machine. The starting mean values for each of the classes are compared with those of true values in Table 1. With these weaker representative points the system started to recognize unknown samples of all the three speakers taken in a random manner from the sample space Ω_X. Though the long and short varieties are pooled together for vowels /I/, /U/, /E/ and /O/, they were given individual reference vectors and weighted vectors computed over the respective set of training samples. Thus in the present experiment $m=6$, $N=3$, $h=1$ for /ə/ and /a :/ and $h=2$ for /I/, /U/, /E/ and /O/. Computing membership values (considering $F_e=1$, $F_d=100$) w.r.t. all the classes, an input utterance is assigned to the kth class ($k=1, 2, \ldots, 6$) associated with maximum μ-value.

For non-supervised learning the decision of the classifier is considered to be final and the parameters of that very recognized class were modified with the addition of a new event before the next input pattern has entered into the system. In supervised learning the decision of the classifier is verified by an

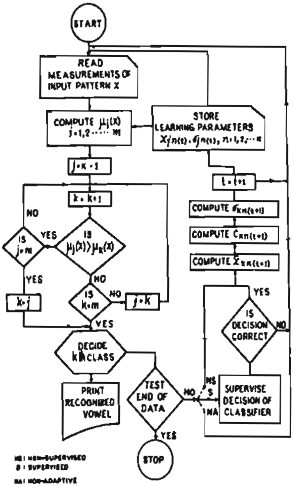


Figure 2. A flow-chart for recognition scheme.

Vowel		Initial			True	
	F_1	F_2	F_3	F_1	F_2	F_3
ə	590	1650	2360	603	1468	2379
a :	690	1280	2220	698	1240	2338
i	330	2280	2988	349	2120	2758
i :	310	2310	3036	335	2286	2853
u	390	960	2680	373	1054	2461
u :	360	830	2632	345	911	2525
e	560	2020	2720	542	1796	2581
e :	460	2260	2910	463	1947	2660
o	550	1160	2718	485	1116	2459
o :	500	990	2822	479	998	2538

Table 1. Comparison of initial and true values of mean vectors.

external supervisor and class parameters are altered only if the classification is found to be correct. Otherwise no alteration of the representative and the weighted vectors is made.

Since the performance of the system in recognizing patterns depends on the sequence of incoming utterances, the experiment was repeated a number of times for the different orders of appearance of the events in sample space.

In the second part of the experiment the initial reference parameters were computed from a training set of 20 samples of all the three speakers selected randomly from each of the classes. This sample size was found to be sufficient in characterizing a class and to provide optimum recognition score (Dutta Majumder *et al.* 1976, Pal and Dutta Majumder 1977 a). The significance of this part was to investigate whether the optimum training set of samples has any impact on the correct rate when the decision is rendered on an adaptive basis.

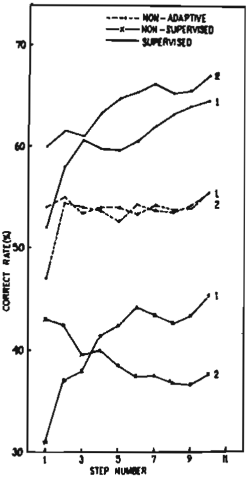
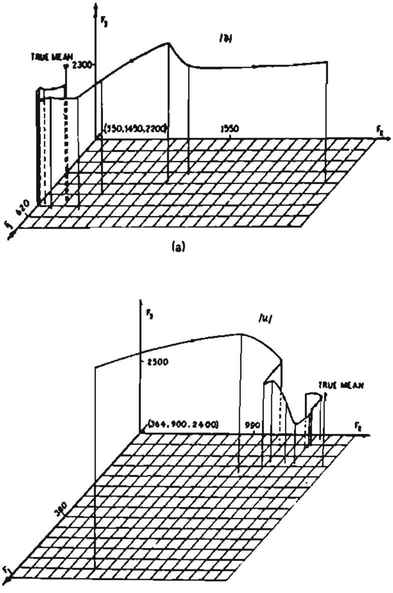


Figure 3. System performance curves.

Figure 3 shows the variation of the cumulative recognition score for two typical instances with successive input patterns where the rate of correct classification after every 100 input samples was noted, and their average result computed at every instant has been plotted. The initial prototype points were estimated with five utterances of a single speaker corresponding to each of the classes. The non-supervised learning algorithm is seen to provide performance inferior to the non-adaptive scheme. The sequence of events happened to be such for the set 2 that the large number of wrong classifications further weakened the already weak representative points. This sequence thus successively worsened the classification situation. But in the case of set 1, the sequence provided a successively better set of input events, which tended to bring the representative points towards the true values and thus the system performance was gradually increasing. A reference to the non-adaptive curve for this set reveals that the second group of 100 events contained markedly better samples than the first group. This obviously improved the representative points significantly. The next groups further helped to keep this trend. This resulted in an upward movement of the curve. Ultimately, the curve tended to approach the classical asymptotic curve for an optimum learning system. For a supervised algorithm, as the name implies, the machine becomes properly acquainted more and more with the unknown patterns as the process of classification continued and the learned parameters for decision gradually approached the true values. The identification efficiency during the system's operation is therefore gradually increased, as is revealed in the corresponding curves for both the sets 1 and 2. This becomes clearer as we enter the pictures in Fig. 4 where the shifting of the mean vectors after every 100 input samples being dealt with are plotted. To restrict the size of the paper, the graphs only for the shorter categories of vowels /∂/ and /u/ were sketched. Three-dimensional effects of the loci are clarified by drawing the front and reverse sides of the curves with solid and broken lines



(a)

(b)

Figure 4. Locus of learned representative vector with progress of learning.

respectively. With initial values of the mean vectors, the movement ultimately approaches the true sample mean, which demonstrated the convergence property of the learning algorithm.

		Recognized as					
		I	E	∂	a :	O	U
Spoken	I	103	29				
		98	34				
	E	8	129	18		11	1
		5	131	19		10	2
	∂		7	26	15	4	
			7	31	10	3	1
	a :			6	56	7	
				8	55	6	
	O		1		1	106	32
			1	5		102	32
	U			1		7	103
				1		4	106

Table 2. Confusion matrix of adaptive vowel recognition when the number of initial training samples is 20. Upper score = non-supervised learning; lower score = supervised learning.

A confusion matrix with an initial learning set of 20 utterances for each vowel is shown in Table 2. The accuracy rates of the individual categories do not differ significantly, whether the system uses an external teacher for supervised learning or not. This should lead to almost equal overall recognition scores for both the systems. In fact, the overall score is found to be identical (≃78%) in both cases. It can therefore be inferred that such an initial training sample size does contain sufficient information about the common properties of a class and provides suitable starting representative vectors of the clusters. The decision taken by the classifier itself, therefore, could be considered almost to be correct. The role of an external supervisor becomes trivial. Once the optimum size of the training set is obtained by the classifier, further increase in the size of the set does not improve the system's performance significantly. The result supports our previous findings (Dutta Majumder *et al.* 1976, Pal and Dutta Majumder 1977 a). Again, the confusion, as expected from our previous impression (Dutta Majumder *et al.* 1976, Dutta Majumder and Pal 1977 b), is seen to be restricted within only two neighbouring classes constituting a vowel triangle.

4. Conclusions

An adaptive learning algorithm from the standpoint of a fuzzy set theory concept is presented and implemented to Telugu vowel sound recognition using a single pattern training procedure. Since the input patterns could not probably be taken serially in a truly random manner, the decision of the classifier and ultimate recognition score are dependent on the sequence of incoming samples. For non-supervised learning, once the patterns are misclassified in a significant proportion, these affect cumulatively the representative vectors, so that the learning process instead of improving may deteriorate the system's performance even w. r. t. a non-adaptive system. This is found to be prominent for weaker initial representative points. The performance of the machine is better for larger initial training samples. Again, if the size of the learning samples reaches an optimum value, the machine with supervised learning did not show any improved result compared to non-supervised learning.

Acknowledgment

The authors wish to acknowledge with thanks the valuable help rendered by Shri N. R. Gańguli, Shri B. Mukherjee, Shri S. Roy, Shri S. Chakraborty and Shri J. Gupta in completing the work.

References

Abramson, N., and Braverman, D., 1962, *I.R.E. Trans. Inf. Theory*, 8, 558.

Chien, Y. T., and Fu, K. S., 1967, *I.E.E.E. Trans. Syst. Man Cybernet.*, 3, 28.

Duda, R. O., and Fossum, H., 1966, *I.E.E.E. Trans. electron. Comput.*, 15, 220.

Dutta Majumder, D., and Datta, A. K., 1969, *J. Inst. electron. telecommun. Engrs*, 15, 233.

Dutta Majumder, D., Datta, A. K., and Pal, S. K., 1976, *J. comput. Soc. India*, 7, 14.

Dutta Majumder, D., and Pál, S. K., 1977 a, *Proceedings of the I.E.E.E. International Conference on Cybernetics and Society*, Washington D.C., 19–21 September (in the press); 1977 b, *Indian J. Tech.* (in the press).

Fu, K. S., 1968, *Sequential Methods in Pattern Recognition and Machine Learning* (London: Academic Press).

Fu, K. S., Chien, Y. T., Nikolic, Z. J., and Wee, W. G., 1966, Tech. Rept., TR-EE66–6, Purdue University, Lafayette, Indiana.

Gaines, B. R., 1975, *Electronics Lett.*, 11, 188.

Ho, Y. C., and Kashyap, R. L., 1965, *I.E.E.E. Trans. electron. Comput.*; 14, 683.

Koford, J. S., and Groner, G. F., 1966, *I.E.E.E. Trans. Inf. Theory*, 12, 42.

Klatt, D. H., and Stevens, K. N., 1973, *I.E.E.E. Trans. Audio Electroacoust.*, 21, 210.

Lindgren, N., 1965, *I.E.E.E. Spectrum*, 2; 1967, *Ibid.*, 4, 75.

Meisel, W. S., 1972, *Computer Oriented Approaches to Pattern Recognition* (New York: Academic Press).

Patric, E. A., and Hancock, J. C., 1966, *I.E.E.E. Trans. Inf. Theory*, 12, 362.

Proc. I.E.E.E., 1976, *Man–Mach. Commun. Voice* (Special Issue), 64, 401.

PAL, S. K., and DUTTA MAJUMDER, D., 1977 a, *I.E.E.E. Trans. Syst. Man Cybernet.*, **7,** 625 ; 1977 b, *Int. J. Systems Sci.*, **9,** 873.

REDDY, D. R., 1975, *Speech Recognition : Invited Papers of the I.E.E.E. Symposium*, 1974, (New York : Academic Press) ; 1976, *Proc. I.E.E.E.*, **64,** 501.

SHARMA, V. V. S., and YEGNANARAYANA, B., 1977, *Proceedings of the All India Interdisciplinary Symposium, Digitals Technique and Pattern Recognition* (Calcutta: Indian Statistical Institute) (in the press).

STALLINGS, W., 1977, *I.E.E.E. Trans. Syst. Man Cybernet.*, **7,** 216.

SEBESTYEN, G. S., 1962, *Decision Making Processes in Pattern Recognition* (New York : Macmillan Co.).

Trans. I.E.E.E. Acoust., Speech Sig. Process, 1975, Special Issue, **23,** 1.

WEE, W. G., and FU, K. S., 1968, *I.E.E.E. Trans. electron. Comput.*, **17,** 178.

WEINSTEIN, C. J., MCCANDLESS, S. S., MONDSHEIN, L. F., and ZUE, V. W., 1975, *I.E.E.E. Trans. Acoust., Speech, Sig. Process.*, **23,** 54.

WOODS, W. A., and MAKHOUL, J., 1974, *Artif. Intell.*, **5,** 73.

ZADEH, L. A., FU, K. S., TANAKA, K., and SHIMURA, M., 1975, *Fuzzy Sets and their Applications to Cognitive and Decision Processes* (New York : Academic Press).

ZADEH, L. A., 1968, *J. math. Analysis Applic.*, **23,** 421.