2nd Hawaii Int'l Conf. on System Sciences,
Honolulu, HI, 1969, pp 341-344

NON-PARAMETRIC UNSUPERVISED LEARNING: IDEAS AND RESULTS*

R. M. Haralick, Univ. of Kansas
Center for Research
Lawrence, Kansas

E. M. Darling, Jr., Natl. Aero. and Space Adm.
Electronics Research Center
Cambridge, Massachusetts

I. Ideas

Let $G = \{g\}$ be measurement space and P be a probability distribution on G. Let $E = \{E_i\}_{i=1}^{N}$ be a given set of subsets of G. We define the characteristic function $\delta: G \to \{-1,1\}^N$ by $\delta(g) = \begin{pmatrix} \delta_1(g) \\ \delta_2(g) \\ \vdots \\ \delta_N(g) \end{pmatrix}$ where $\left.\begin{array}{l} \delta_i(g) = 1 \text{ if and only if } g \in E_i \\ \quad = -1 \text{ otherwise.} \end{array}\right\}$, $i = 1, 2, \ldots, N$.

Let $\langle g_1, g_2, \ldots, g_t, \ldots \rangle$ be a sequence of elements sampled from G according to probability distribution P. On the basis of this sequence and Q_o we describe first a predictive and then a non-predictive learning procedure which defines a sequence of partitions $\langle H^1, H^2, \ldots, H^n, \ldots \rangle$ such that as n gets large, the cells of the partition H^n tend to become similarity sets.

By "similarity sets" we mean subsets of measurement space having highly correlated or similar elements. The basic idea behind the procedure is that if it is possible to say that an element $g \in G$ belongs to the similarity set H_k, then it also should be possible to accurately describe that particular g from the general characterization of H_k.

This can occur in our proceedure because each q_{ij} measures the total relationship between δ_i and η_j. This relationship is a function of a_{ij}, b_{ij}, c_{ij}, d_{ij}:

a_{ij} measures the extent to which $+\delta_i$ and $+\eta_j$ are both +1.
b_{ij} measures the extent to which $-\delta_i$ and $-\eta_j$ are both +1.
c_{ij} measures the extent to which $-\delta_i$ and $+\eta_j$ are both +1.
d_{ij} measures the extent to which $+\delta_i$ and $-\eta_j$ are both +1.

Let $\{a_{ij}^o, b_{ij}^o, c_{ij}^o, d_{ij}^o \mid i = 1, \ldots, N; j = 1, \ldots, K\}$ be a set of parameters whose values are chosen arbitrarily. We will later impose constraints on these parameters as well as introduce other parameters.

Define $q_{ij}^o = \dfrac{a_{ij}^o b_{ij}^o - c_{ij}^o d_{ij}^o}{a_{ij}^o b_{ij}^o + c_{ij}^o d_{ij}^o}$ and the N x K matrix $Q_o = (q_{ij}^o)$. The set H^o is a partition over G and will be defined as a function of the Q_o matrix. $H^o = \{H_j^o\}_{j=1}^{2^K}$, $H_j^o = \{g \mid \eta(g) = \text{Sgn } Q_o' \delta(g) \text{ and } 1 + \sum_{k=1}^{K} (\eta_k(g)+1)2^{k-2} = j\}$, $j = 1, \ldots, 2^K$ where $\text{Sgn} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{pmatrix}$ $\left.\begin{array}{l} \text{and } s_i = 1 \text{ if and only if } v_i > 0 \\ \quad = -1 \text{ otherwise} \end{array}\right\}$.

In what follows the N-tuple $\delta(g)$ can be considered as a detailed description of g, in fact, it might even be so detailed that there is a one-to-one correspondence between g and $\delta(g)$. The function $\eta: \{-1,1\}^N \to \{-1,1\}^K$ is defined by $\eta(g) = \begin{pmatrix} \eta_1(g) \\ \eta_2(g) \\ \vdots \\ \eta_K(g) \end{pmatrix} = \text{Sgn } Q'\delta(g)$, and $\eta(g)$ can be considered as the characterization of the similarity set to which g belongs. The N-tuple Θ acts as a threshold point for the function $\hat{\delta}: \{-1,1\}^K \to \{-1,1\}^N$ defined by $\hat{\delta}(g) = \text{Sgn}[Q\eta(g) - \Theta]$. We may consider $\hat{\delta}(g)$ the description of g obtained on the basis of the knowledge that g is in the similarity set characterized by $\eta(g)$. The N-tuple γ indicates the bias: the component γ_i is the relative percentage that $\hat{\delta}_i(g) = +1$ and $\delta_i(g) = -1$ minus the relative percentage that $\hat{\delta}_i(g) = -1$ and $\delta_i(g) = +1$. ϵ^* is the reinforcement parameter. r_{1j} is the relative percentage that any component $\delta_i(g) = \hat{\delta}_i(g)$ when $\eta_j(g) = +1$.
r_{2j} is the relative percentage that any component $\delta_i(g) \neq \hat{\delta}_i(g)$ when $\eta_j(g) = +1$.
r_{3j} is the relative percentage that any component $\delta_i(g) = \hat{\delta}_i(g)$ when $\eta_j(g) = -1$.
r_{4j} is the relative percentage that any component $\delta_i(g) \neq \hat{\delta}_i(g)$ when $\eta_j(g) = -1$.

*Sponsored by the Department of Defense; Contract DAAK02-68-C-0089.

In addition, the parameters a_{ij}, b_{ij}, c_{ij}, d_{ij} satisfy the constraints for $i = 1,\ldots,N$, $j = 1,\ldots, K$:

(1) $a_{ij} + b_{ij} + c_{ij} + d_{ij} = 1$

(2) $a_{ij} + c_{ij} = a_{nj} + c_{nj}$, $n = 1,\ldots, N$

(3) $a_{ij} + d_{ij} = a_{in} + d_{in}$, $n = 1,\ldots, K$

(4) $0 \leq a_{ij}, b_{ij}, c_{ij}, d_{ij} \leq 1$

The predictive procedure is indicated in flow-chart one. Note that the symbol "←" means replacement or substitution. For instance, the statement "a ← a+1" means replace a by a+1.

The non-predictive procedure eliminates the computation of $\hat{\delta}$, γ, and Θ, and the reinforcement always acts to increase the association between $\delta(g)$ and $\eta(g)$. The only constraint which the parameters a_{ij}, b_{ij}, c_{ij}, d_{ij} satisfy is that they remain non-negative. The definitions for q_{ij}, η, and H also differ:

$$q_{ij} = \frac{a_{ij}b_{ij}}{c_{ij}d_{ij}} \text{ and } \eta(g) = \begin{pmatrix} \eta_1(g) \\ \eta_2(g) \\ \vdots \\ \eta_K(g) \end{pmatrix} \quad \text{where } \eta_j(g) = 1 \text{ if and only if } \prod_{i=1}^{N} (q_{ij})^{\delta_i(g)} \geq \prod_{i=1}^{N} (q_{in})^{\delta_i(g)} \text{ for all } n$$

$$= 0 \text{ otherwise}$$

$$H = \{H_i\}_{i=1}^{K}, \; H_i = \{g \mid \eta_i(g) = 1\}$$

The non-predictive procedure is indicated in flow-chart two.

II. Results

The predictive learning procedure was tried for an artificial data set in which there were four normally distributed clusters. The parameters a_{ij}, b_{ij}, c_{ij}, and d_{ij} were initially chosen from a uniform distribution and then made to satisfy the constraints. The reinforcement parameter ϵ^* was set equal to 5×10^{-4} and K = 3, N = 8. The learning curve in Figure 1 indicates the probability of classifying an element g in the category to which it belongs. In addition, the non-predictive learning procedure was used for a real set of agricultural radar imagery. There was an 85 - 90% correspondence between the similarity sets and the known categories.

The non-predictive learning procedure was tried for an artificial data set in which there were two normally distributed clusters. The parameters a_{ij}, b_{ij}, c_{ij}, and d_{ij} were intially chosen to equal .25. The reinforcement parameter ϵ^* was set equal to 10^{-3} and K = 2, N = 10. The learning curve in Figure 2 indicates the probability of classifying an element g in the category to which it ideally belongs.

The learning curves for the predictive and non-predictive procedure show quick learning in the first few hundred iterations and then a general leveling off. Sometimes after this leveling, additional learning will occur.

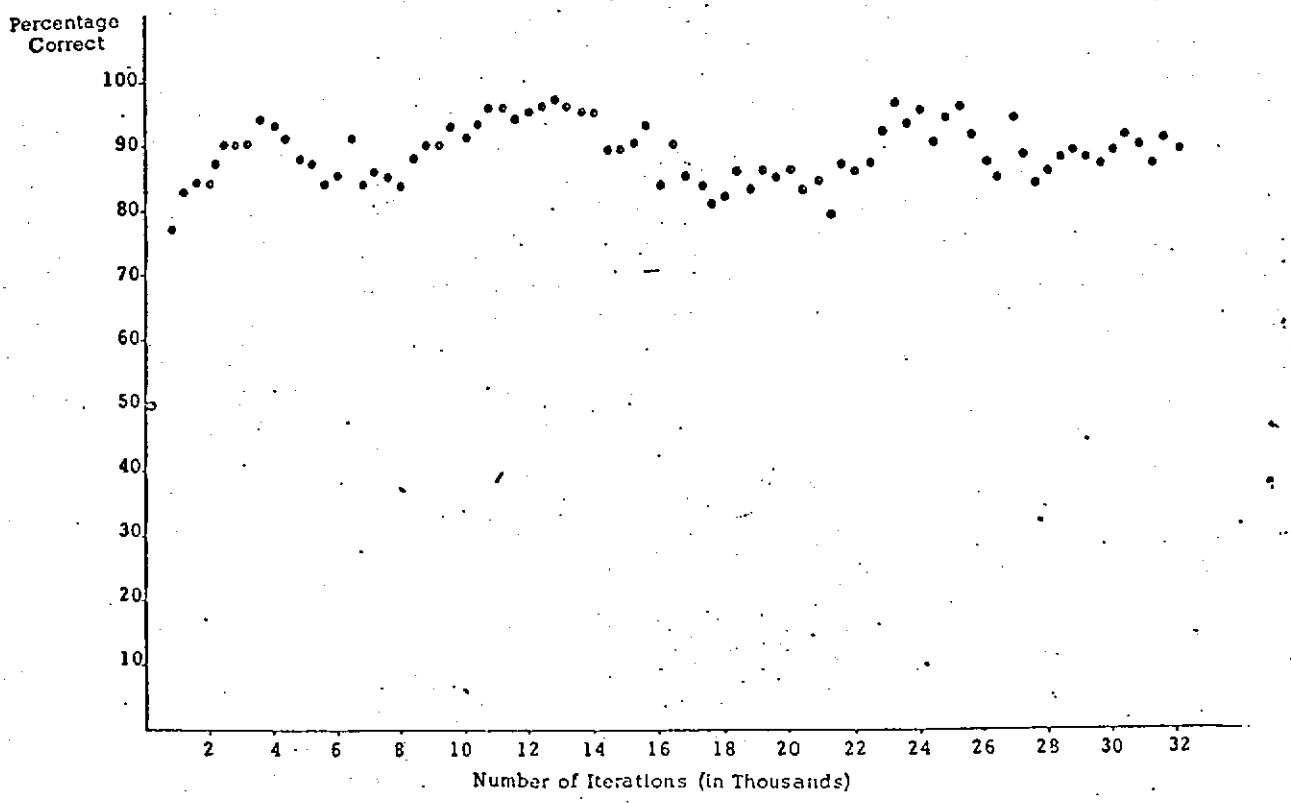
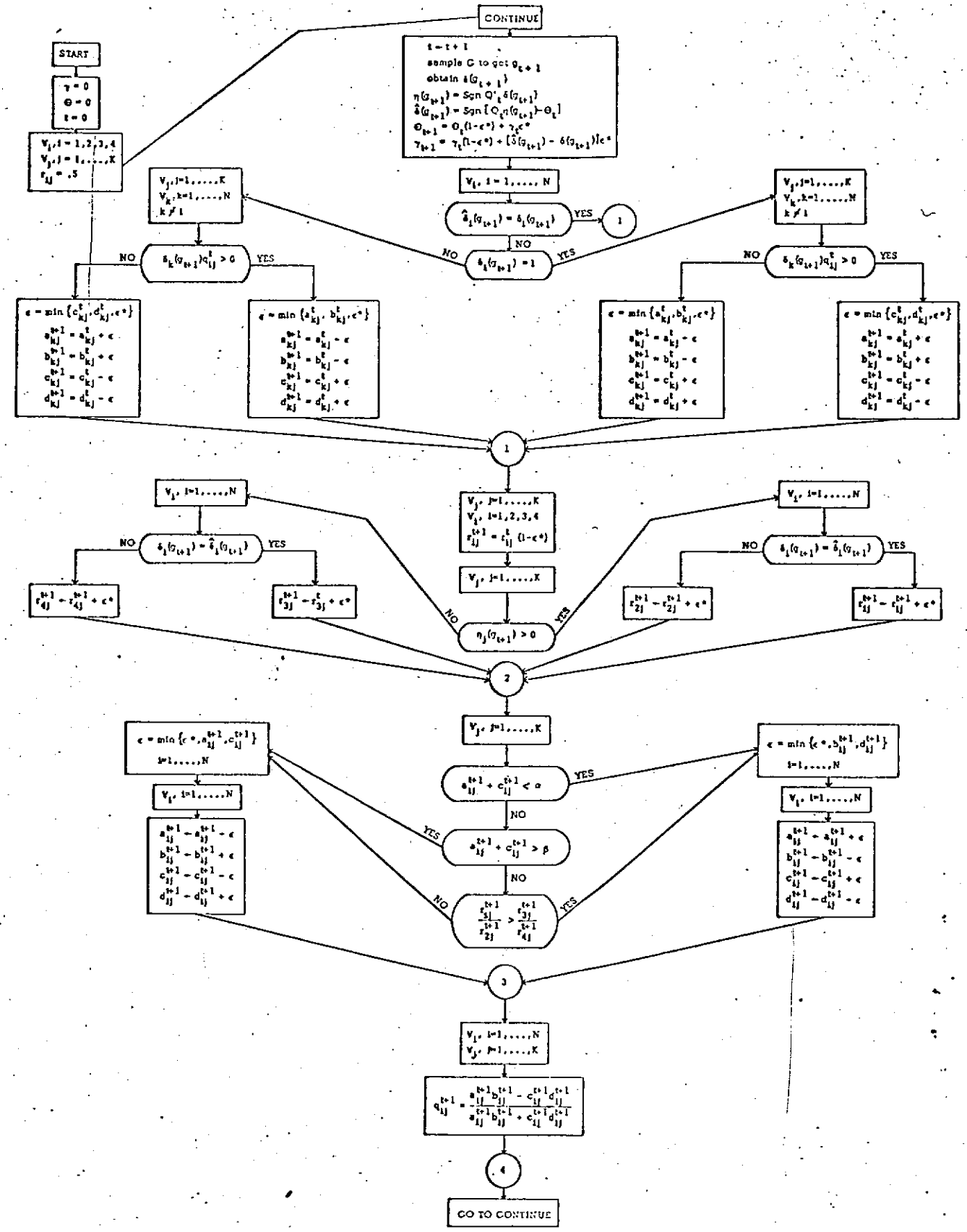
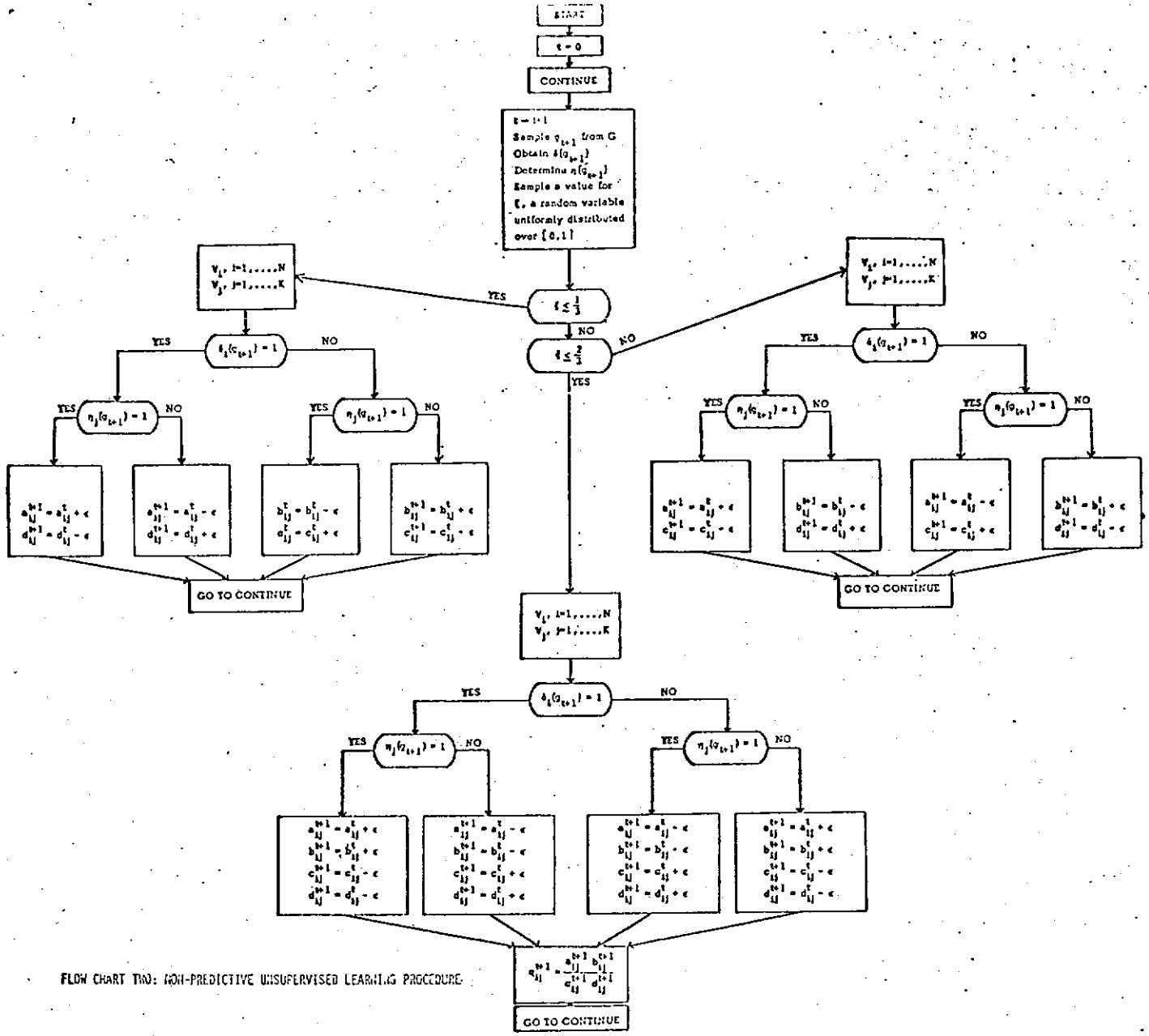


Figure 1. Learning Curve for Unsupervised non-parametric Predictive Procedure.



FLOW CHART ONE: - PREDICTIVE UNSUPERVISED LEARNING PROCEDURE



FLOW CHART TWO: NON-PREDICTIVE UNSUPERVISED LEARNING PROCEDURE

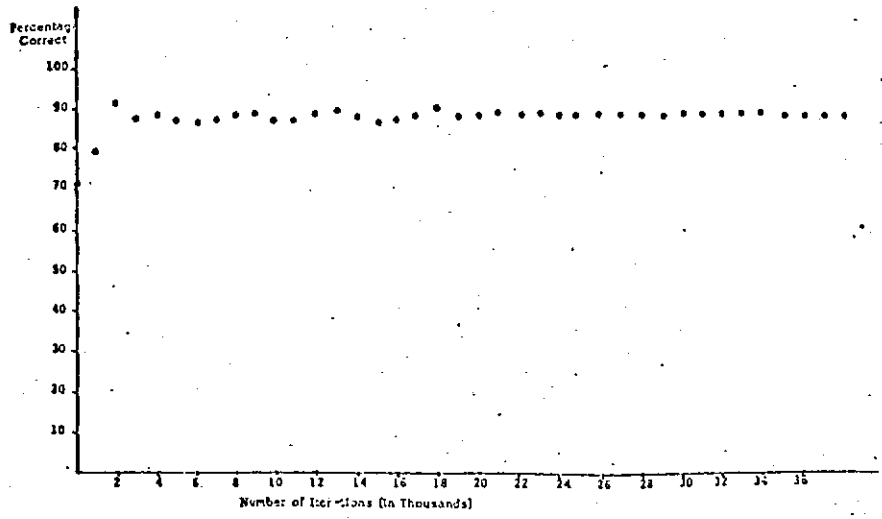


Figure 2. Learning Curve for Unsupervised Non-Parametric Non-Predictive Procedure

*This research was in part sponsored by Advanced Research Projects Agency (ARPA), Department of Defense, Work Order No. 1079, Monitored for ARPA by USAETL Contract No. DAAK 02-68-C-0089.