7.9-10092
CR 151544

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
MANNED SPACECRAFT CENTER HOUSTON, TEXAS

ISODATA: THRESHOLDS FOR SPLITTING CLUSTERS

by

(E79-10092) ISODATA: THRESHOLDS FOR SPLITTING CLUSTERS (Lockheed Electronics Co.) 38 p HC A03/MF A01 CSCL 02C

N79-15370

Unclas
G3/43 00092

Houston Aerospace Systems Division
LOCKHEED ELECTRONICS COMPANY, INC.
A Subsidiary of Lockheed Aircraft Corporation

LEC/HASD No. 640-TR-058
LEC Job Order 74-524
A.D. No. 63-0117-4524

ISODATA: THRESHOLDS FOR SPLITTING CLUSTERS

by

Edwin P. F. Kan

Prepared for:

DATA APPLICATIONS BRANCH

EARTH OBSERVATIONS DIVISION

National Aeronautics and Space Administration
Manned Spacecraft Center
Houston, Texas

January 1972

EOD2011

LEC/HASD No. 640-TR-058
LEC Job Order 74-524
A.D. No. 63-0117-4524

ISODATA: THRESHOLDS FOR SPLITTING CLUSTERS

January 1972

Prepared by: Edwin P. F. Kan
Data Techniques Section

Approved by: C. E. Clouse, Supervisor
Data Techniques Section

Approved by: S. B. Cousin, Manager
Data Applications Department

DATA APPLICATIONS DEPARTMENT

LOCKHEED ELECTRONICS COMPANY, INC.
HOUSTON AEROSPACE SYSTEMS DIVISION
HOUSTON, TEXAS 77058

ABSTRACT

Thresholds for the parameter AD* which decides on the splitting of a cluster in the ISODATA program are discussed. For the univariate case, 0.84 is established as a sound threshold, after testing on some typical distributions, simple as well as composite, and evaluating the probability of misclassification. Extension to the multivariate case leads to the empirical value of $(N-0.16)/\sqrt{N}$, where N denotes the dimension of the vector space. A critical examination of the values of AD, especially for large N, results in the conclusion that AD is not a very effective measure for the present purpose.

*AD is the acronym for "average distance". Its exact definition is found in equation (1) of Section 1.

TABLE OF CONTENTS

1. INTRODUCTION

The iterative unsupervised clustering ISODATA program proposed by Ball and Hall [1,2,3] has been very well received. Application of such technique for the eventual goal of classification of multivariate statistical data has resulted in various levels of success by different organizations in diversified disciplines such as in remote sensing. One major difficulty in using ISODATA* is the lack of universally acknowledged thresholds for various parameters that appear in the program. Although experience can suggest to the programmer the appropriate values, it is certainly necessary to establish analytically some sound values for these thresholds. It is the intent of this paper to do exactly this in regard to the splitting of clusters.

Summarizing the idea of ISODATA without going to its mathematical details, the following gives a verbal account of the program. The goal of the procedure is to separate all the statistical data into classes, or clusters, each having its average point (i.e., mean) as its representative point. At each intermediate step, new assignment of points is conducted by grouping or splitting the old clusters. The decisions on the assignment of points to different clusters, on the splitting and on the combining of clusters depend on some distance function and parameters derived from these distance measures. The program is started by arbitrary (or some preferred) initialization of cluster centers, and ends by noting the invariance of clusters with respect to additional iterations.

---

*The present discussion is primarily on ISODATA-POINTS [3].

A measure of the tightness of packing of points $\{X_k\}_{k=1}^n$, $X_k = (X_{k1}, X_{k2}, \cdots, X_{kN})^T \in R^N$, in a cluster C is the 'average distance' AD defined by

$$AD \triangleq \frac{1}{n} \sum_{k=1}^{n} d(X_k, \mu) \quad , \tag{1}$$

where $\mu \equiv (\mu_1, \mu_2, \cdots \mu_N)^T$ is the representative point of C , i.e., the mean

$$\mu_j \triangleq \frac{1}{n} \sum_{k=1}^{n} X_{kj} \quad , \tag{2}$$

and $d(\cdot,\cdot) : R^N \times R^N \to R$ is a distance function

$$d^2(X,Y) \triangleq \sum_{i=1}^{N} w_i (X_i - Y_i)^2 \quad , \tag{3}$$

$w_i \in R^+$. A commonly* used set of weights $\{w_i\}$ is

$$w_i = \frac{1}{\sigma_i^2} \quad , \quad \sigma_i \neq 0 \quad , \tag{4}$$

where σ_i^2 is assumed to be nonzero and denotes the variance of the cluster in the *ith* coordinate axis

---

*Too many people in various disciplines have used this measure. Thus, no tracing to its original user is possible.

$$\sigma_i^2 \triangleq \frac{1}{n} \sum_{k=1}^{n} (X_{ki} - \mu_i)^2 \quad . \tag{5}$$

This $\{w_i\}$ is a reasonable choice because of the standardization or 'sphericalization' of the cluster C. Let T be the threshold. The rule to decide on the splitting of a cluster is:

0. If $AD \geq T$, split the cluster.
1. Otherwise, no splitting.

The following further examines this parameter AD. The bound $\sqrt{N}$ is established, i.e., AD is proved to be $\leq \sqrt{N}$. The values of AD for univariate normal, triangular, trapezoidal, rectangular and 'bi-spiked' distributions are computed. Furthermore, different composite univariate distributions, each made up of two normal distributions with various variances and *a priori* probabilities, are studied. The probabilities of misclassification using the Bayes decision rule for the composite distributions are also computed in the case when a cluster is decided to be split into two, each assuming a normal distribution, and that points are to be reassigned. These considerations lead to the conclusion that the value 0.84 of AD is a sound choice for the threshold to split a cluster in ISODATA for the univariate case. While the value $(N-0.16)/\sqrt{N}$ of AD is extrapolated as an empirical threshold for the multivariate case, its effectiveness as a critical, discriminative parameter is considerably reduced when it is noticed that AD approaches its bound $\sqrt{N}$ fairly rapidly with N. In fact, for N large, the AD for any reasonably smooth distribution is shown to approach $\sqrt{N}$, which makes it impossible to decide whether the distribution is simple or composite, i.e., which makes it impossible to decide whether to split the cluster or not.

2. SOME PROPERTIES OF AD

The intimacy between the tightness of a cluster and its AD has been suggested in the previous section. To further study this association, some properties of AD are obtained here.

2.1 Upper Bound of AD

<u>Fact</u>: With the definition of AD through (1)-(5), $AD \le \sqrt{N}$.

Proof:

$$AD^2 = \frac{1}{n^2}\left[\sum_{k=1}^{n} d(X_k,\mu)\right]^2$$

$$\le \frac{1}{n}\sum_{k=1}^{n} d^2(X_k,\mu) \quad , \quad \text{by Lemma (see below)}$$

$$= \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}(X_{ki}-\mu_i)^2 \quad , \quad \sigma_i \ne 0$$

$$= \sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left\{\frac{1}{n}\sum_{k=1}^{n}(X_{ki}-\mu_i)^2\right\}$$

$$= \sum_{i=1}^{N}\frac{1}{\sigma_i^2}\sigma_i^2$$

$$= N \quad .$$

Lemma

$$\sum_{k=1}^{N} x_k^2 \geq \frac{1}{N}\left(\sum_{k=1}^{N} x_k\right)^2 , \qquad x_k \in R .$$

Proof: (By mathematical induction).

It suffices to prove the Lemma for the case when $x_k \in R^+$.

(a) $N = 2$:

$$\text{L.H.S.} = x_1^2 + x_2^2$$

$$= \frac{1}{2}\left(2x_1^2 + 2x_2^2\right) . \tag{6}$$

Since $x_1^2 + x_2^2 \geq 2x_1x_2$,

$$(6) \geq \frac{1}{2}\left[\left(x_1^2 + x_2^2\right) + 2x_1x_2\right]$$

$$= \frac{1}{2}(x_1 + x_2)^2$$

$$= \text{R.H.S.}$$

(b) Assume the Lemma is true for $N = M$, check it for $M+1$:

$$\sum_{k=1}^{M+1} x_k^2 = \sum_{k=1}^{M} x_k^2 + x_{M+1}^2 \geq \frac{1}{M} y^2 + x_{M+1}^2 \quad , \qquad (7)$$

where $y \equiv \sum_{k=1}^{M} x_k$, and the Lemma is applied for $N = M$.

$$(7) = \frac{1}{M+1} y^2 + \frac{1}{M(M+1)} y^2 + x_{M+1}^2$$

$$= \frac{1}{M+1} \left\{ y^2 + \left[\frac{1}{M} y^2 + M x_{M+1}^2 \right] + x_{M+1}^2 \right\} \quad . \qquad (8)$$

As $\left[\frac{1}{M} y^2 + M x_{M+1}^2\right] \geq 2 y x_{M+1}$,

$$(8) \geq \frac{1}{M+1} \left(y^2 + 2 y x_{M+1} + x_{M+1}^2 \right) = \frac{1}{M+1} (y + x_{M+1})^2$$

That is,

$$\sum_{k=1}^{M+1} x_k^2 \geq \frac{1}{M+1} \left(\sum_{k=1}^{M+1} x_k \right)^2 \quad .$$

Observations

1. The equality in the Lemma holds when $|x_i| = |x_j|$, for all $i \neq j$.

2. Thus, the upper bound $\sqrt{N}$ is attained by AD when $d(X_i,\mu) = d(X_j,\mu)$ for all $i \neq j$.

2.2 AD For Some Simple Univariate Distributions

Assumption 1: N is assumed to be 1 in this subsection through Section 4. Extension to the case $N \neq 1$ is relegated to Section 5.

Assumption 2: Enough data points are assumed to be available such that the histograms of these data points will reproduce the assumed distribution.

Notations: In the following, $p(\cdot)$, μ, σ , AD will denote the probability density function, mean, standard deviation and the average distance (defined through (1)-(5)), respectively.

(a) Normal distribution $N(0,\hat{\sigma}^2)$

$$p(x) = \frac{1}{\sqrt{2\pi}\hat{\sigma}} e^{-x^2/2\hat{\sigma}^2}$$

$$\mu = 0$$

$$\sigma = \hat{\sigma}$$

$$
\begin{aligned}
AD &= \int_{-\infty}^{\infty} \frac{|x|}{\sigma}\, p(x)\, dx \\
&= \frac{2}{\sqrt{2\pi}\hat{\sigma}} \int_{0}^{\infty} \frac{x}{\sigma}\, e^{-x^2/2\hat{\sigma}^2}\, dx \\
&= \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-r}\, dr \quad , \quad r \equiv x^2/2\hat{\sigma}^2 \\
&= \frac{2}{\sqrt{2\pi}} \\
&= 0.80 \quad . \qquad (9)
\end{aligned}
$$

(b) Triangular, trapezoidal, rectangular and 'bi-spiked' distributions.

Since the derivation of results for these distributions are straightforward as in (a), the results are only tabulated in Table 1. Refer to Figure 1 for the shapes of these distributions.

Observation: It is obvious that the more widespread the distribution is, in the sense of its span relative to its standard deviation, the larger is AD. Thus, in the 'bi-spiked' case where the density has an extreme allocation, AD is unity, the upper bound given in Section 2.1.

3. AD FOR SOME COMPOSITE UNIVARIATE NORMAL DISTRIBUTIONS

This section addresses to the decision of the splitting of a cluster by studying its AD. Assuming that samples from two distributions, whether they be close or far apart from each other, have been initially taken to belong to the same cluster, i.e. to a composite distribution, the question is: How will the value of AD provide the information as to whether or not it is advisable to split the cluster into two? The following will study a few combinations of two normal distributions $p_1(\cdot)$ and $p_2(\cdot)$ with various variances and *a priori* probabilities π_1 and π_2, i.e., $p(\cdot) = \pi_1 p_1(\cdot) + \pi_2 p_2(\cdot)$. See Figure 2.

Case (a)

$$\pi_1 = \pi_2 = 1/2 \quad , \quad p_1 \sim N(0,1) \quad , \quad p_2 \sim N(-\Delta,1) \; :$$

$$p(x) = \frac{1}{2}\, p_1(x) + \frac{1}{2}\, p_2(x) = \frac{1}{2\sqrt{2\pi}} \left[e^{-x^2/2} + e^{-(x+\Delta)^2/2} \right] .$$

It is simple to show that

$$\mu = 0$$

$$\sigma^2 = (4 + \Delta^2)/4$$

$$\mathrm{AD} = \frac{1}{\sigma} \left\{ \frac{2}{\sqrt{2\pi}}\, e^{-\Delta^2/8} + \frac{\Delta}{2}\, [1 - 2P_N(-\Delta/2)] \right\} \tag{10}$$

where

$$P_N(\alpha) = \int_{-\infty}^{\alpha} p_N(x)\, dx \quad , \quad p_N \sim N(0,1) \quad . \qquad (11)$$

Values of these parameters for various Δ are tabulated in Table 2(a) together with the shape of the composite distributions.

Case (b)

$$\pi_1 = \pi_2 = 1/2 \quad , \quad p_1 \sim N(0,2) \quad , \quad p_2 \sim N(-\Delta,1) :$$

$$p(x) = \frac{1}{2\sqrt{2\pi}} \left[\frac{1}{\sqrt{2}}\, e^{-x^2/4} + e^{-(x+\Delta)^2/2} \right] \quad .$$

It can be similarly shown that

$$\mu = -\Delta/2$$

$$\sigma^2 = (6 + \Delta^2)/4$$

$$AD = \left[\frac{1}{2}\, AD_1 + \frac{1}{2}\, AD_2 \right] \Big/ \sigma \quad ,$$

where

$$AD_1 = \int_{-\infty}^{\infty} |x - \mu|\, p_1(x)\, dx$$

$$= \mu \left[2P_N(\mu/\sqrt{2}) - 1 \right] + \frac{4}{\sqrt{4\pi}}\, e^{-\mu^2/4} \quad ,$$

and

$$AD_2 = \int_{-\infty}^{\infty} |x - \mu| p_2(x)\, dx$$

$$= (\mu + \Delta)[2P_N(\mu + \Delta) - 1] + \frac{2}{\sqrt{2\pi}} e^{-(\mu+\Delta)^2/2} , \qquad (12)$$

where $P_N(\cdot)$ is given by (11). Values of these parameters for various Δ are again tabulated in Table 2(b).

Case (c)

$\pi_1 = 2/3$, $\pi_2 = 1/3$, $p_1 \sim N(0,2)$,

$p_2 \sim N(-\Delta,1)$.

Similarly,

$$\mu = -\Delta/3$$

$$\sigma^2 = (15 + 2\Delta^2)/9$$

$$AD = \left[\frac{2}{3} AD_1 + \frac{1}{3} AD_2\right] \Big/ \sigma \qquad (13)$$

where AD_1 and AD_2 are given in (12). Table 2(c) shows the values of these parameters for different Δ .

Case (d)

$\pi_1 = 3/4 \quad , \quad \pi_2 = 1/4 \quad , \quad p_1 \sim N(0,2) \quad ,$

$p_2 \sim N(-\Delta,1)$

Similarly,

$$\mu = -\Delta/4$$

$$\sigma^2 = (28 + 3\Delta^2)/16$$

$$AD = \left[\frac{3}{4} AD_1 + \frac{1}{4} AD_2\right] \Big/ \sigma \quad , \tag{14}$$

where AD_1 and AD_2 are given in (12). Table 2(d) shows the value of these parameters for different Δ .

Observations

1. Looking at the shape of the composite distribution versus the accompanying value of AD, it can be concluded that when $AD \geq 0.84$, the distributions discussed so far have two distinct "humps". This threshold value 0.84 of AD suggests, then, to be a sound choice to decide whether a cluster should be split or not.

2. In case when a cluster is split, the examples above indicate that a good universal choice of the two new cluster centers is at $\pm\sigma$ from the old cluster center, i.e. μ . This is due to the fact that the points $\mu\pm\sigma$ are quite close to the means of $p_1(\cdot)$ and $p_2(\cdot)$. Indeed, this conclusion conforms with common practice.

3. The following section discussing errors of misclassification will further warrant this value 0.84 of AD.

4. PROBABILITY OF MISCLASSIFICATION

This section investigates the following dilemma. In the case when a composite distribution is made up of two normal distributions with their means at a distance Δ apart, one might be faced with the dilemma that even if these two distributions are identified (i.e., even if decision is made to split the cluster) the reassignment of points (from the old cluster) to their respective distributions (i.e., new clusters) might not be easy. Or, the assignment might involve too much error. If the probability of error of misclassification becomes too high, it might be more justified to consider that only one cluster exists (with a distorted normal distribution) than to consider the co-existence of two clusters (each coming from a normal distribution).

The following will calculate the probability of misclassification P_E that arises when the points are reassigned according to the Bayes decision rule maximizing *a posteriori* probability. This is a familiar technique to the communication engineers (see ref. [4]), and is summarized as follows:

<u>Bayes Rule</u>: Assign a point x to the distribution $p_1(\cdot)$ if $\pi_1 p_1(x) > \pi_2 p_2(x)$. Otherwise, assign it to the distribution $p_2(\cdot)$.

Referring to Figure 2, the Bayes Rule says that any point x with $x > \theta$ will be assigned to distribution $p_1(\cdot)$, and $x \leq \theta$ assigned to distribution $p_2(\cdot)$, where θ is such that

$$p_1(-\theta) = p_2(\Delta - \theta) \quad . \tag{15}$$

Thus, the error of misclassification P_E is

$$P_E = \pi_1 \int_{-\infty}^{-\theta} p_1(x)\, dx + \pi_2 \int_{\Delta-\theta}^{\infty} p_2(x)\, dx \quad . \tag{16}$$

Straightforward calculation will show for the four cases studied in Section 3:

<u>Case (a)</u>

$$\theta = \Delta/2$$

$$P_E = \frac{1}{2} P_N(-\theta) \quad , \tag{17}$$

where $P_N(\cdot)$ is given in (11).

<u>Case (b)</u>

$$\theta = 2\Delta - \sqrt{2\Delta^2 + 4\phi} \quad , \quad \phi = \ln \sqrt{2}$$

$$P_E = \frac{1}{2\sqrt{2}} P_N(-\theta/\sqrt{2}) + \frac{1}{2} [1 - P_N(\Delta - \theta)] \quad . \tag{18}$$

<u>Case (c)</u>

$$\theta = 2\Delta - \sqrt{2\Delta^2 - 4\phi} \quad , \quad \phi = \ln \sqrt{2}$$

$$P_E = \frac{2}{3\sqrt{2}} P_N(-\theta/\sqrt{2}) + \frac{1}{3} [1 - P_N(\Delta - \theta)] \quad . \tag{19}$$

Case (d)

$$\theta = 2\Delta - \sqrt{2\Delta^2 - 4\psi} \quad , \quad \psi = \ln(3/\sqrt{2})$$

$$P_E = \frac{3}{4\sqrt{2}} P_N(-\theta/\sqrt{2}) + \frac{1}{4}[1 - P_N(\Delta - \theta)] \quad . \qquad (20)$$

Values of θ and P_E for different Δ are tabulated in Table 2(a)-(d).

Observations

1. First, it is noticed [5] that in the classification of, say, agricultural data, by remote sensing techniques, a 90% correct classification is considered to be good under the existing state-of-the-art. In other words, an error of up to 10% is permissable.

2. Examination of Table 2 reveals that if a composite distribution with $AD \geq 0.84$ is split into two normal distributions, the error of misclassification P_E in the process of reassigning the points of the old cluster to the two new clusters will be less than 10%. This means that splitting of the cluster is justified. Otherwise, if $AD < 0.84$, it is more advisable to retain the old cluster and assume that it originates from a distorted normal distribution.

3. The threshold 0.84 of AD is thus further justified.

5. EXTENSION TO MULTIVARIATE CASE

It is readily seen that extension of the results obtained in the previous sections to the multivariate case is nontrivial if not impossible. However, the evaluation of AD for the simplest multivariate normal distribution is still possible when mutual uncorrelation (and hence independence) is assumed of the components of the N-vector. The value of AD obtained for this distribution will certainly serve as a guideline for the values of AD for other simple or composite distributions. Furthermore, it is recalled that the bound on AD has been shown to be $\sqrt{N}$. Thus, an empirical threshold for AD can be extrapolated.

5.1 AD For Multivariate Normal Distribution

Assume the following probability density:

$$p(x) = \frac{1}{(2\pi)^{N/2} \prod_{i=1}^{N} \sigma_i} \exp\left(-\sum_{i=1}^{N} x_i^2/2\sigma_i^2\right) . \tag{21}$$

<u>N = 1</u>: As shown in Section 2.2,

$$\begin{aligned} AD &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-r}\, dr \\ &= 0.80 \quad . \end{aligned} \tag{22}$$

$\underline{N = 2}$:

$$AD = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2}\right) \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{x_1^2}{2\sigma_1^2} - \frac{x_2^2}{2\sigma_2^2}\right) dx_1\ dx_2$$

$$= \frac{1}{2\pi}\int_0^{\infty} re^{-r^2/2}(2\pi r)\ dr \quad , \quad r \equiv \left(\sum_{i=1}^{2} x_i^2/\sigma_i^2\right)^{1/2}$$

$$= \int_0^{\infty} r^2 e^{-r^2/2}\ dr$$

$$= 1.25 \quad . \tag{23}$$

$\underline{N = 3}$: Similarly,

$$AD = \frac{1}{(2\pi)^{3/2}}\int_0^{\infty} re^{-r^2/2}(4\pi r^2)\ dr \quad , \quad r \equiv \left(\sum_{i=1}^{3} x_i^2/\sigma_i^2\right)^{1/2}$$

$$= \sqrt{\frac{2}{\pi}}\int_0^{\infty} r^3 e^{-r^2/2}\ dr$$

$$= 1.60 \quad . \tag{24}$$

$\underline{N = 4}$: Similarly,

$$AD = \frac{1}{(2\pi)^2}\int_0^\infty re^{-r^2/2}(2\pi^2 r^3)\, dr \quad , \quad r \equiv \left(\sum_{i=1}^{4} x_i^2/\sigma_i^2\right)^{1/2}$$

$$= \frac{1}{2}\int_0^\infty r^4 e^{-r^2/2}\, dr$$

$$= 1.88 \quad . \tag{25}$$

Curve (a) in Figure 3 shows how $AD/\sqrt{N}$ varies with N for this normal distribution.

5.2 Empirical Threshold

Since 0.84 has been shown to be a sound threshold for the univariate case $N = 1$, an empirical value of AD can be extrapolated for the multivariate case. The empirical formula

$$\text{Threshold} = (N - 0.16)/\sqrt{N} \tag{26}$$

appears to be a reasonable rule as plotted in Figure 3, curve (b). Typically then, the threshold for AD when $N = 3$ is $(3 - 0.16)/\sqrt{3} = 1.64$.

5.3 Critical Observation

From Figure 3, it is observed that the value of AD for a simple multivariate distribution (as given by (21)) approaches its limit $\sqrt{N}$ reasonably fast with N . In fact, examining the forms of (22)-(25), the factor r^N in the integral is readily recognized to be that factor which drives the value of AD for any reasonably smooth distribution (which possesses the first absolute moment) to its bound $\sqrt{N}$. This is by virtue of the fact that, for N large, the integral

$$2\int_0^\infty r^N e^{-r^2/2}\, dr \tag{27}$$

equals $2\int_0^\infty r(r^{N-1}e^{-r^2/2})\, dr \equiv \int_{-\infty}^\infty |x|\hat{p}(x)\, dx$, when $\hat{p}(s)$ will approximate the 'bi-spiked' distribution as shown in Figure 1(e), with $\Delta = 1$. Consequently, (27) and thus AD approaches its bound.

CRITICISM: With the above observation, it is asserted that the parameter AD will lose its effectiveness as a discriminative parameter to decide on the splitting of a cluster when N is large. This is because the AD of any (reasonable) distribution approaches $\sqrt{N}$ and makes it impossible to detect the structure of the distribution from the value of AD.

6. CONCLUSION

The parameter AD as used in the ISODATA program is critically examined. Thresholds of AD to decide on the splitting of clusters are obtained. For the univariate case, 0.84 is established as a sound choice, after examining several simple as well as composite distributions and also after investigating the probability of misclassification when points have to be reassigned to the newly identified clusters. For the multivariate case, the empirical threshold $(N-0.16)/\sqrt{N}$ is extrapolated. A final criticism on AD is that AD would lose its effectiveness as a discriminative measure for the present purpose when N is large.

ACKNOWLEDGMENT

The author would like to thank T. C. Minter for his interest and suggestive discussions that lead to Section 4 of the paper. He is also indebted to W. A. Holley for the practical insight into the problem.

REFERENCES

1. G. H. Ball and D. J. Hall, "ISODATA, an Iterative Method of Multivariate Analysis and Pattern Classification," Proceedings of the International Communication Conference, Philadelphia, Pa., June 1966.

2. G. H. Ball and D. J. Hall, "A Clustering Technique for Summarizing Multivariate Data," Behavorial Sciences, Vol. 12, No. 2, pp. 153-155, March 1967.

3. G. H. Ball and D. J. Hall, "ISODATA, a Novel Method of Data Analysis and Pattern Classification," Stanford Res. Inst., Menlo Park, Calif., Tech. Rep. AD 699616, April 1965.

4. M. Schwartz, W. Bennett and S. Stein, Communication Systems and Techniques, McGraw Hill Book Co., New York, N. Y., 1966.

5. K. S. Fu, D. A. Langrebe and T. L. Phillips, "Information Processing of Remotely Sensed Agricultural Data," Proc. IEEE, Vol. 57, No. 4, pp. 639-653, April 1969.

TABLE CAPTIONS

1. Tabulation of the probability density, mean, standard deviation and AD for the various distributions of Figure 1.

2. Tabulation with respect to Δ of the mean, standard deviation, AD, θ and P_E of the composite distribution $p(\cdot) = \pi_1 p_1(\cdot) + \pi_2 p_2(\cdot)$ with

   (a) $\pi_1 = \pi_2 = 1/2$, $p_1 \sim N(0,1)$, $p_2 \sim N(-\Delta,1)$

   (b) $\pi_1 = \pi_2 = 1/2$, $p_1 \sim N(0,2)$, $p_2 \sim N(-\Delta,1)$

   (c) $\pi_1 = 2/3$, $\pi_2 = 1/3$, $p_1 \sim N(0,2)$, $p_2 \sim N(-\Delta,1)$

and (d) $\pi_1 = 3/4$, $\pi_2 = 1/4$, $p_1 \sim N(0,2)$, $p_2 \sim N(-\Delta,1)$.

FIGURE CAPTIONS

1. Some simple distributions: (a) Normal $N(0,\hat{\sigma}^2)$, (b) Triangular with spread 2Δ, (c) Trapezoidal with spread 4Δ, (d) Rectangular with spread 2Δ and (e) 'Bi-spiked' with equal weights spread 2Δ apart.

2. A composite distribution $p(\cdot)$ made up of two normal distributions $p_1(\cdot)$ and $p_2(\cdot)$ with *a priori* probabilities π_1 and π_2 .

3. Variation of $AD/\sqrt{N}$ with N for (a) Normal distribution, (b) The suggested threshold: $AD = (N - 0.16)/\sqrt{N}$.

	Normal	Triangular	Trapezoidal	Rectangular	'Bi-Spiked'
$p(x)$	$\frac{1}{\sqrt{2\pi}\hat{\sigma}} e^{-x^2/2\hat{\sigma}^2}$	$\frac{1}{\Delta}\left(1 - \frac{\lvert x\rvert}{\Delta}\right)$, $\lvert x\rvert \le \Delta$; $0,\ \lvert x\rvert > \Delta$	$\frac{1}{3\Delta},\ \lvert x\rvert \le \Delta$; $\frac{1}{3\Delta}\left(2 - \frac{\lvert x\rvert}{\Delta}\right)$, $\Delta < \lvert x\rvert \le 2\Delta$; $0,\ \lvert x\rvert > 2\Delta$	$\frac{1}{2\Delta},\ \lvert x\rvert \le \Delta$; $0,\ \lvert x\rvert > \Delta$	$\frac{1}{2}\delta(x - \Delta)$ $+ \frac{1}{2}\delta(x + \Delta)$
μ	0	0	0	0	0
σ	$\hat{\sigma}$	$\frac{1}{\sqrt{6}}\Delta$	$\sqrt{\frac{5}{6}}\Delta$	$\frac{1}{\sqrt{3}}\Delta$	Δ
AD	0.80	0.82	0.85	0.87	1

Table 1

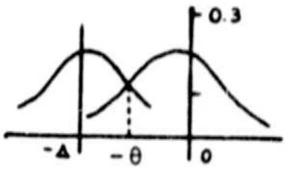
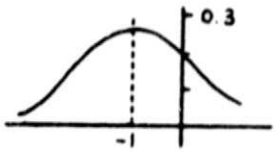
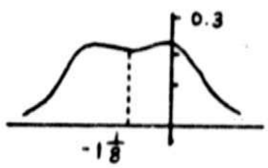
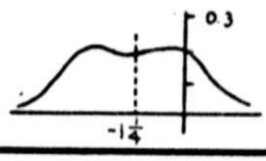
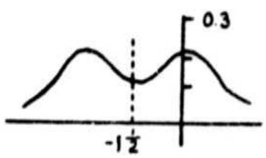
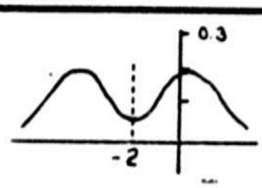
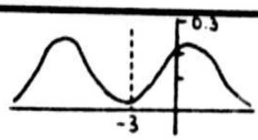
Δ	$p(x)$	μ	σ	AD	$-\theta$	P_E
		Section 3			Section 4	
2		-1	1.41	0.82	-1	0.16
$2\frac{1}{4}$		$-1\frac{1}{8}$	1.51	0.83	$-1\frac{1}{8}$	0.13
$2\frac{1}{2}$		$-1\frac{1}{4}$	1.60	0.84	$-1\frac{1}{4}$	0.11
3		$-1\frac{1}{2}$	1.80	0.87	$-1\frac{1}{2}$	0.07
4		-2	2.24	0.90	-2	0.02
6		-3	3.16	0.95	-3	0.00

Table 2(a)

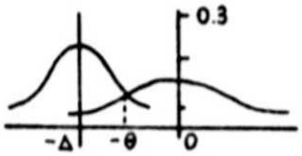
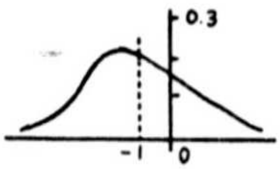
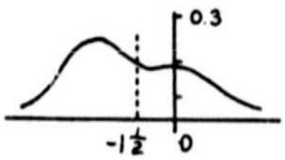
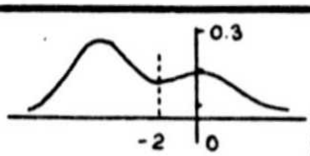
Δ	p(x)	μ	σ	AD	$-\theta$	P_E
		Section 3			Section 4	
2		-1	1.58	0.81	-0.94	0.162
3		$-1\ \frac{1}{2}$	1.94	0.84	-1.60	0.087
4		-2	2.34	0.88	-2.22	0.039

Table 2(b)

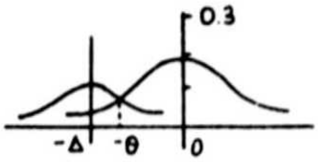
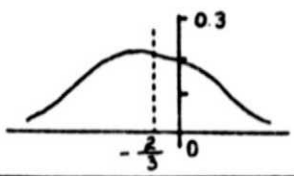
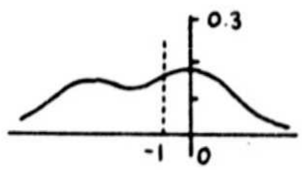
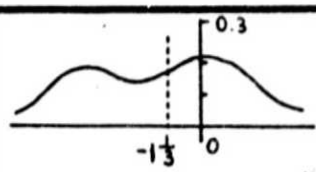
Δ	p(x)	μ	σ	AD	-θ	P_E
		Section 3			Section 4	
2		$-\frac{2}{3}$	1.60	0.82	-1.43	0.169
3		-1	1.91	0.84	-1.92	0.086
4		$-1\frac{1}{3}$	2.29	0.84	-2.46	0.042

Table 2(c)

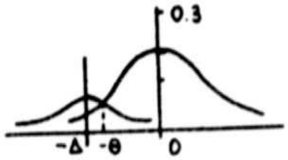
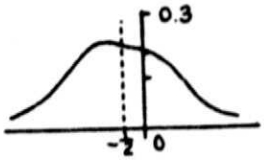
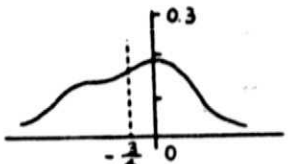
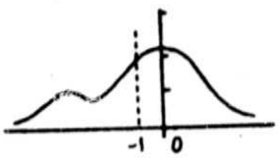
Δ	p(x)	μ	σ	AD	-θ	P_E
		Section 3			Section 4	
2		$-\frac{1}{2}$	1.58	0.83	-1.76	0.159
3		$-\frac{3}{4}$	1.85	0.83	-2.12	0.083
4		-1	2.18	0.88	-2.62	0.038

Table 2(d)

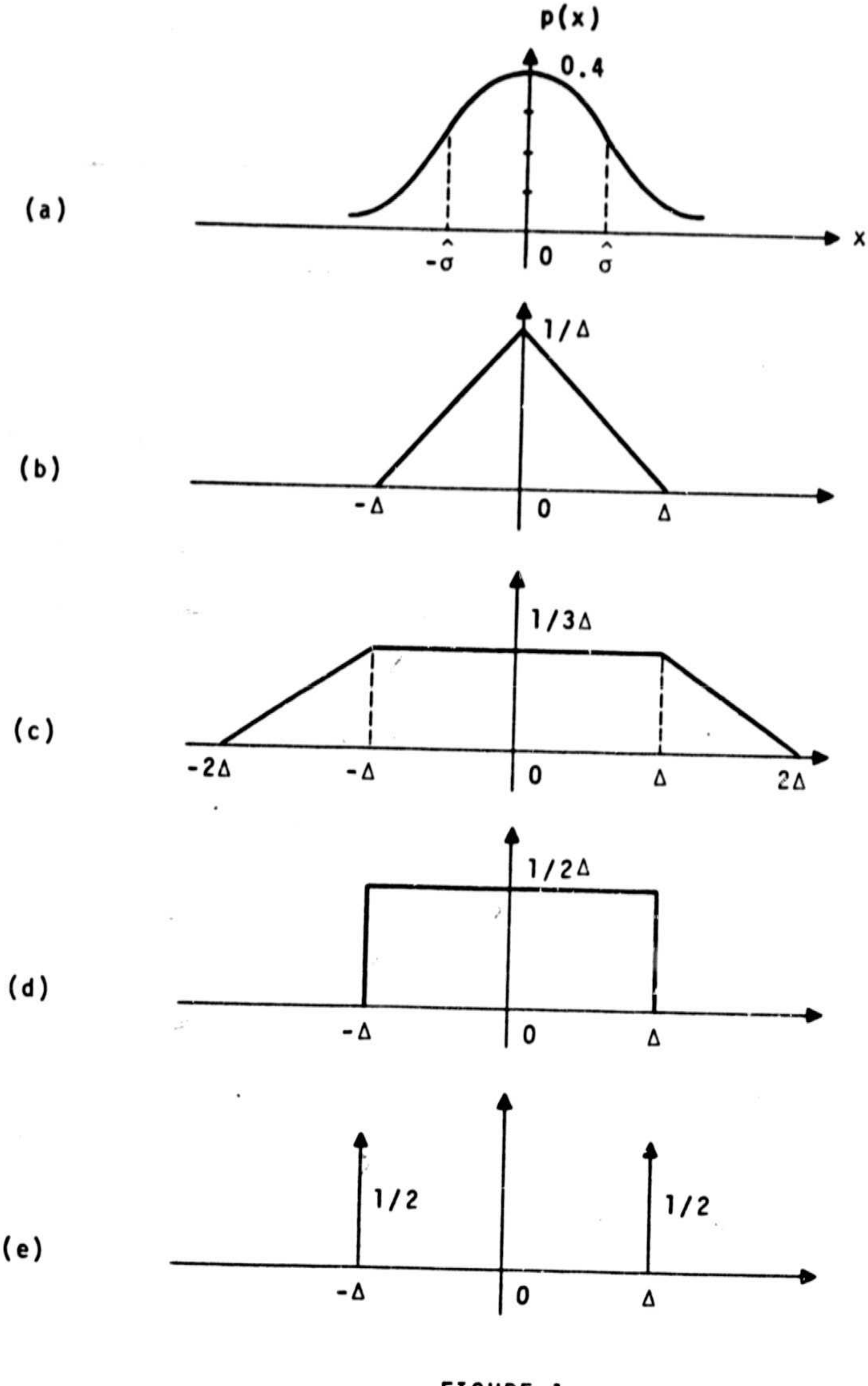


FIGURE 1

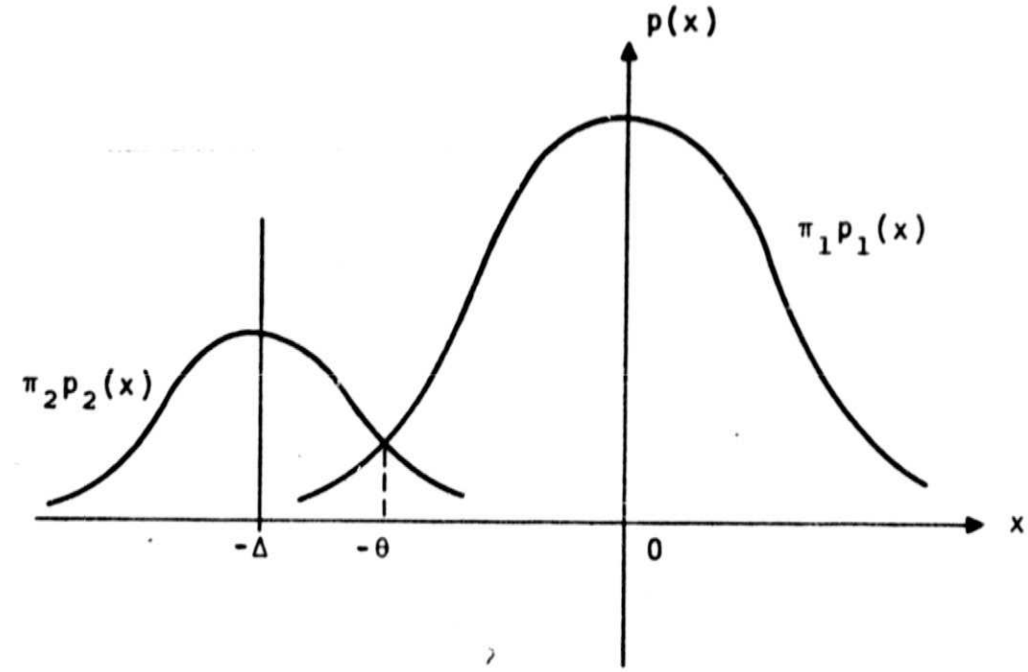


FIGURE 2

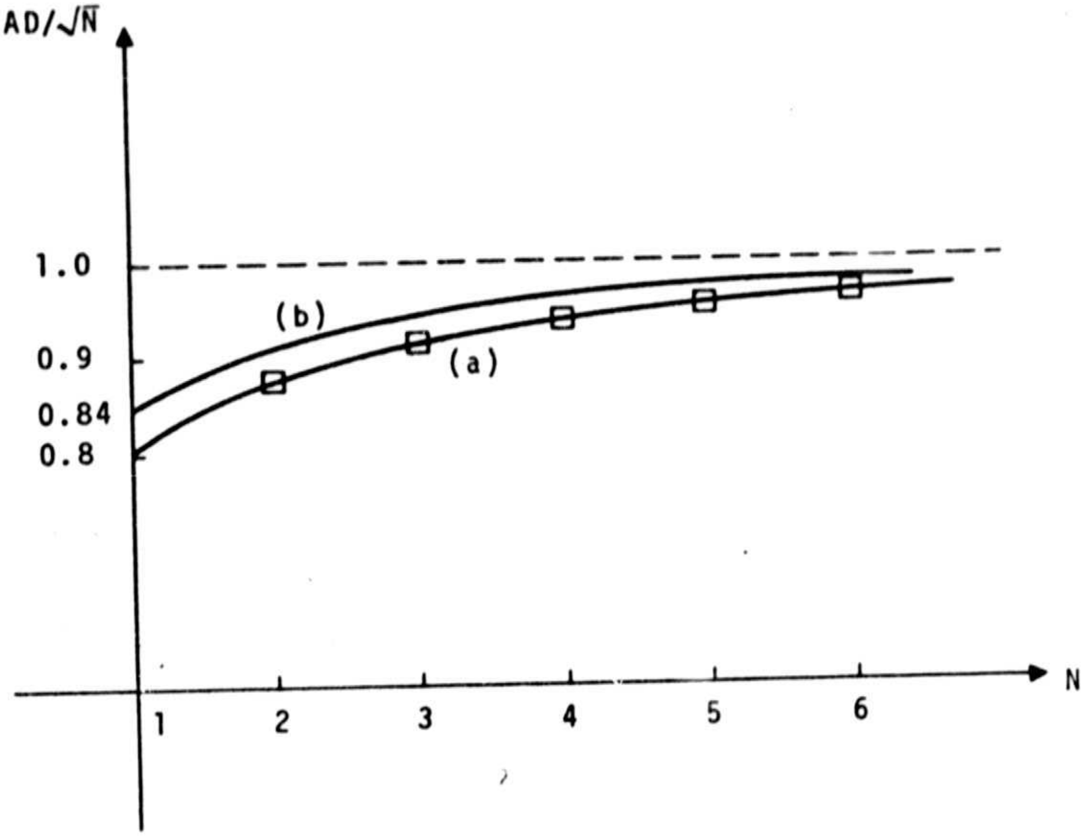


FIGURE 3