Instead of probing at the inputs and outputs of the first candidate, we can replace it with a circuit board that is known to be good. Then, we apply the same test to the primary inputs and compare the output vector after the replacement with the old one. If the two vectors are identical, then the original board was not faulty and diagnosis proceeds with the remaining candidates. If they are different and the new one is correct, then the original board was faulty and diagnosis can be terminated. Otherwise, the original board was faulty and there are still unknown faults. The diagnosis continues on the remaining candidates reordered according to the new symptom.

The result of applying a different test to the device can also be used to reorder candidates. For example, we can move candidates connecting to corroborations under the new test to the end of the candidate list and move candidates connecting to violations to the front.

The ordered candidate list is sufficient for selecting the best action among the same type but is not sufficient for selecting actions of different types. To select actions among different types a thorough analysis of probabilities and action costs is necessary [2. 4].

References

[1] B. Chandrasekaran and R. Milne, "Reasoning about structure, behavior and function," *ACM SIGART Newsletter*, vol. 93, Jul. 1985.

[2] J. S. Chen, "A Probabilistic Theory of Model Based Diagnosis," *Ph.D. Thesis*, Department of Computer Science, University at Buffalo, Buffalo. NY, 1992, *Technical Report 92-04*.

[3] J. S. Chen and S. N. Srihari, "Candidate ordering and elimination in model-based fault diagnosis," in *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 1363 1368, Morgan Kaufmann, 1989.

[4] J. S. Chen and S. N. Srihari, "Action selection in interactive model-based diagnosis," in *Proceedings of the Eighth IEEE Conference on Artificial Intelligence Applications*, pp. 67 73, 1992.

[5] H. C. A. Dale, "Fault-finding in electronic equipment," *Ergonomics*, vol. 1, pp. 356–385, 1957.

[6] R. Davis, "Diagnostic reasoning based on structure and behavior," *Artificial Intelligence*, vol. 24, no. 3, pp. 347–410, 1984.

[7] R. Davis and W. Hamscher, "Model-based reasoning: Troubleshooting," in H. Shrobe, editor. *Exploring Artificial Intelligence: Survey Talks from the National Conferences on Artificial Intelligence*, ch. 8, pp. 297–346, Morgan Kaufmann, San Mateo, CA, 1988.

[8] R. Davis, H. Shrobe, W. Hamscher. K. Wieckert, M. Shirley and S. Polit. "Diagnosis based on description of structure and function," in *Proceedings of the Second National Conference on Artificial Intelligence*, pp. 137–142. Morgan Kaufmann, 1982.

[9] J. de Kleer, "An assumption-based TMS," *Artificial Intelligence*, vol. 28, pp. 127–162, 1986.

[10] J. de Kleer and B. C. Williams, "Diagnosing multiple faults," *Artificial Intelligence*, vol. 32, no. 1, pp. 97–130, 1987.

[11] M. R. Genesereth, "The use of design description in automated diagnosis," *Artificial Intelligence*, vol. 24, no. 3, pp. 411–436. 1984.

[12] R. M. Hunt and W. B. Rouse, "A fuzzy rule-based model of human problem solving," *IEEE Trans. on Syst., Man, and Cybern.*, SMC-14, no. 1, pp. 112–120, Jan./Feb. 1984.

[13] J. Rasmussen and A. Jensen, "Mental procedures in real-life tasks: A case study of electronic trouble shooting," *Ergonomics*, vol. 17, no. 3, pp. 293–307, 1974.

[14] J. A. Reggia, D. S. Nau and P. Y. Wang, "Diagnostic expert system based on a set covering model," *International Journal of Man-Machine Studies*, vol. 19, pp. 437–460, 1983.

[15] J. A. Reggia, D. S. Nau and P. Y. Wang, "A formal model of diagnostic inference. I. Problem formulation and decomposition," *Information Sciences*. vol. 37, pp. 227 256, 1985.

[16] R. Reiter, "A theory of diagnosis from first principles," *Artificial Intelligence*, vol. 32, no. 1, pp. 57–95, 1987.

[17] W. B. Rouse, "Human problem solving performance in a fault diagnosis task," *IEEE Trans. on Syst., Man, and Cybern.*, SMC-8, no. 4, pp. 258–271, Apr. 1978.

[18] W. B. Rouse, "A model of human decision making in a fault diagnosis task," *IEEE Trans. on Syst., Man, and Cybern.*, SMC-8, no. 5, pp. 357–361. May 1978.

[19] W. B. Rouse, "Problem solving performance of first semester maintenance trainees in two fault diagnosis tasks," *Human Factors*, vol. 21, no. 5, pp. 611–618, Oct. 1979.

[20] W. B. Rouse, "Problem solving performance of maintenance trainees in a fault diagnosis task," *Human Factors*, vol. 21, no. 2, pp. 195–203, Apr. 1979.

[21] W. B. Rouse. "Models of natural intelligence in fault diagnosis tasks: Implications for training and aiding of maintenance personnel," in *Proceedings of the Joint Services Workshop on Artificial Intelligence in Maintenance*, pp. 193–212, Jun. 1984.

[22] W. B. Rouse, S. H. Rouse and S. J. Pellegrino, "A rule-based model of human problem solving performance in fault diagnosis tasks," *IEEE Trans. on Syst., Man, and Cybern.*, SMC-10, no. 7, pp. 366–376, Jul. 1980.

[23] S. C. Shapiro, S. N. Srihari, M. R. Taie and J. Geller, VMES: "A network-based versatile maintenance expert system," in *Proceedings of the 1st International Conference on Applications of AI to Engineering Problems*, pp. 925 936. Springer-Verlag, Apr. 1986.

[24] M. R. Taie, J. Geller, S. N. Srihari and S. C. Shapiro. "Knowledge based modeling of circuit boards," in *Proceedings of 1987 Annual Reliability and Maintainability Symposium*, pp. 422–427, Jan. 1987.

[25] Z. Xiang, "Multi-level Model-based Diagnostic Reasoning," *Ph.D. Thesis*, Department of Computer Science, University at Buffalo, Buffalo, NY 14260, Aug. 1988, *Technical Report 88-17*.

[26] Z. Xiang and S. N. Srihari, "A strategy for diagnosis based on empirical and model knowledge," in *Proceedings of the Sixth International Workshop on Expert Systems and Their Applications*, pp. 835–848, Avignon, France, Apr. 1986.

Finding a Subset of Representative Points in a Data Set

D. Chaudhuri, C. A. Murthy, and B. B. Chaudhuri

***Abstract*—This correspondence deals with the problem of finding the representative points from a data set $\subseteq \Re^2$. Two algorithms are stated. One of the algorithms can find the *local best representative* or *seed* points. The extension of these algorithms for three or more dimensions is also discussed. Experimental results on synthetic and real life data are provided which manifest the utility of these algorithms.**

I. Introduction

Consider a set of objects represented as a point data in $\Re^2$ feature space. Given n data, we address the problem of selecting a small subset of $n_0 \ll n$ data that *faithfully* represents the spatial organization of the original data. The solution to this problem can find applications in data compression, data clustering [8], [14], pattern classification, as well as statistical parameter estimation [9], [14]. For example, in clustering, many algorithms start with a few *seed* points, where each *seed* point represents the core of one cluster. The minimum distance classifier or the K-nearest neighbor classifier needs the *seed* points as the best patterns representing the classes. Quite often, the *best representative point* may be the *mode* of the pattern set. *Density* and *mode* estimation are two classical problems in statistics. In many situations, the problem of finding the best

Manuscript received June 22, 1992; revised August 26, 1993.

The authors are with the Electronics and Communication Sciences Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India.

IEEE Log Number 9401467.

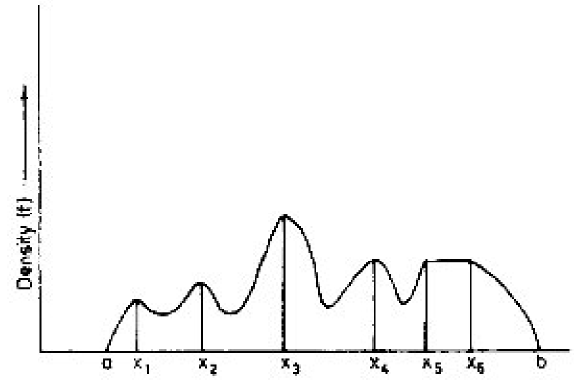


Fig. 1. A density function f on a set $[a, b]$.

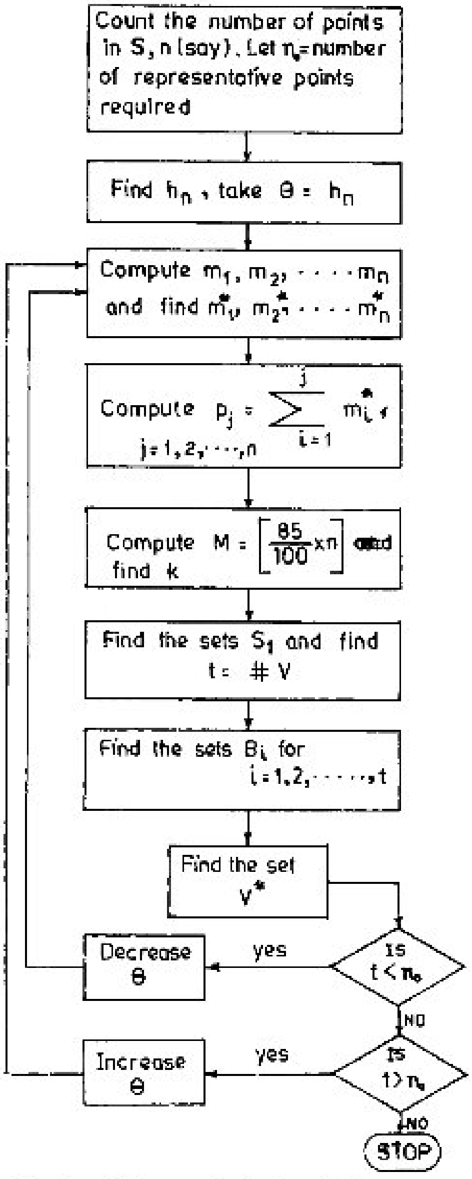

Count the number of points in S, n (say). Let n_0 = number of representative points required
Find h_n, take $\theta = h_n$
Compute $m_1, m_2, \cdots m_n$ and find $m_1^*, m_2^*, \cdots m_n^*$
Compute $p_j = \sum_{i=1}^{j} m_i^*$, $j = 1, 2, \cdots, n$
Compute $M = \left[\frac{85}{100} \times n\right]$ and find k
Find the sets S_1 and find $t = \# V$
Find the sets B_i for $i = 1, 2, \cdots, t$
Find the set V^*
Is $t < n_0$
yes
Decrease θ
NO
Is $t > n_0$
yes
Increase θ
NO
STOP

Fig. 2. A flowchart of the proposed algorithm AL-2.

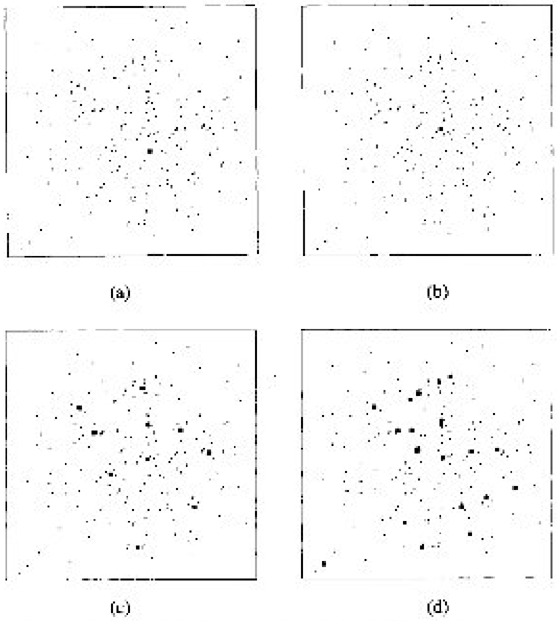


Fig. 3. A triangular distribution data of size 200. (a) Local best representative point. (b) Cluster center by applying C-means algorithm ($C = 1$). (c) Almost 5% representative points. (d) Almost 10% representative points.

representative points may be considered as the generalization of mode estimation and seed point detection problem.

In his classical work [1], Parzen showed how to construct a family of consistent and asymptotically unbiased estimates of probability density function and its mode. The work was extended by Cacoullos [2], among others [9]. In the literature of cluster analysis, there exist several approaches of *seed* point estimation [8]. Macqueen [3] chooses the first k data units in data set as the initial *seed* points. Forgy [4] takes any desired partition of the dta units into k mutually exclusive groups and computes the group centroids as *seed* points. Astrahan [5] computes the density for each data unit as the number of other data units within some specified distance, orders the data units by density, and chooses the one with the highest density as the first *seed* point. The subsequent *seed* points are chosen in order of decreasing density, subject to the stipulation that each new *seed* point is at least a minimum distance away from all other previously chosen *seed* points. A simpler approach was suggested by Ball and Hall [6]. Here, the overall mean vector of data set is considered as the first *seed* point. The subsequent *seed* points are selected by examining the data units in their input sequence and accepting any data unit which is at least some specified distance, say d, from all previously chosen *seed* points. The process is continued until k *seed* points are accumulated or the data set is exhausted. Ling [11] suggests that each (k, r)-cluster has the property that its elements are within a distance r of at least k other elements of the same cluster, and the entire set can be marked by a chain of links each of length less than or equal to r. But there are no guidelines about k and r.

Given n data, the problem is that of finding a small subset of $n_0 \ll n$ data that provides good representation of the original data. Let a representative point p represent r data in its neighborhood. Intuitively, p should be such that the sum of its distances from these r data is minimum. But an exact algorithm, which gives the optimal representative point set, is computationally very expensive. See Section VI for details.

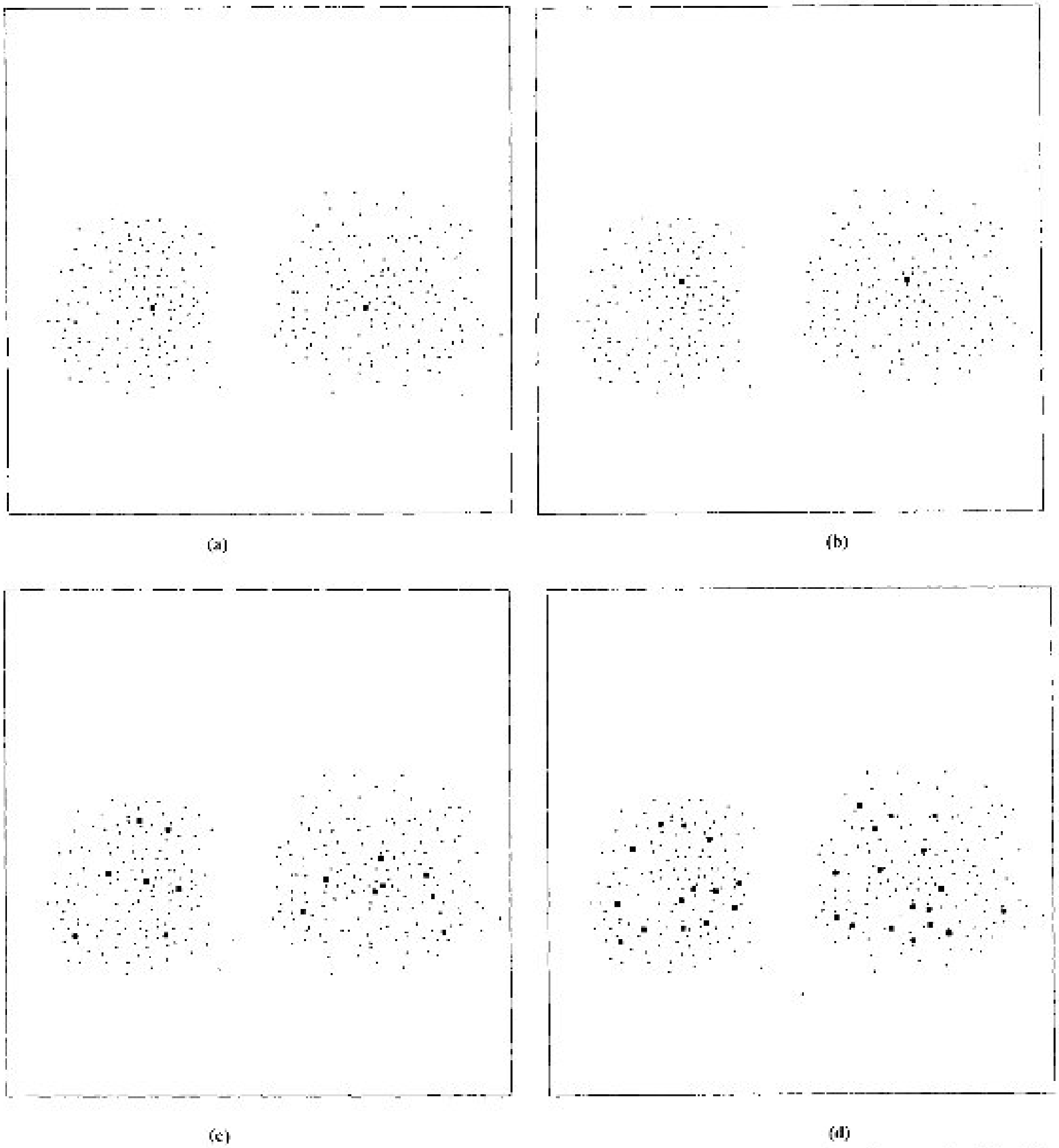


Fig. 4. Synthetic data of two clusters of size 322. (a) Local best representative point. (b) Cluster center by applying C-means algorithm ($C = 2$). (c) Almost 5% representative points. (d) Almost 10% representative points.

Our approach for finding the local best representative points in plane is dependent upon the local densities of the points. The density of any point x is assumed to be the number of data points present in an open disc of radius θ around x. Minimal spanning tree (MST) of the data points is used to decide the value of θ. Euclidean interpoint distance is taken as the edge weight of the MST. The square root of the average edge weight h_n is taken to be equal to θ. A previous work by Murthy [12] shows the importance of h_n in the set estimation problem.

It may be noted that all the *high density* points may not give *good* representative points because many high density points may remain in a small locality and the representatives do not cover the entire data set. Thus, the distance between any two *good representative points* should be significant. The h_n mentioned above has been used to fix the distance between any pair of representative points. Two algorithms are presented in this correspondence for finding the local best representative point set. The first algorithm can be used when n_0 is fixed *a priori*, while the second algorithm can be used when n_0 is supplied. The utility of these algorithms is successfully tested on various artificial data sets and also on a real life speech sound data set.

The organization of the correspondence is as follows. Section II provides a few definitions of representative points. Section III gives the description of two algorithms when the data is from $\Re^2$. Section IV provides the results of the two algorithms. Section V deals with the extension of these algorithms to higher dimensions and provides the results of applying them on a speech recognition problem.

II. A Few Definitions

This section contains a few definitions related to locally best representative points and the basic setup under which the algorithms

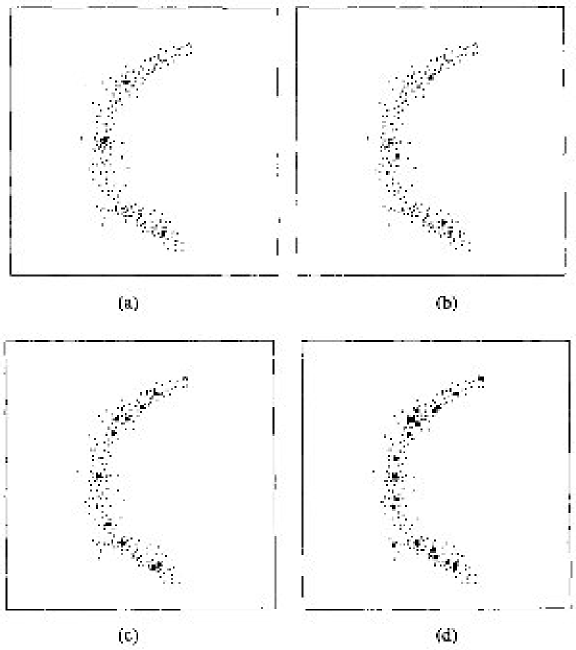


Fig. 5. C-shaped synthetic data of size 214. (a) Local best representative points. (b) Cluster center by applying C-means algorithm ($C = 3$). (c) Almost 5% representative points. (d) Almost 10% representative points.

in the following sections have been stated. An example is given below to explain the definitions.

Example: A density function f on a set $[a, b]$ has been shown in Fig. 1.

1) Note that for every point x in $A = \{x_1, x_2, x_3, x_4\}$, there exists an $\epsilon_x > 0$ such that

$$f(x) > f(y) \qquad \forall\, y \in (x - \epsilon_x, x + \epsilon_x);$$
$$y \neq x, x \in A.$$

We denote every $x \in A$ as *local best representative point* of f.

2) Note also that $f(x_3) > f(x) \qquad \forall\, x \in [a, b]$ and $x \neq x_3$. We denote x_3 as the *global best representative point.*

3) Note that $f(x) = f(y) \qquad \forall\, x, y \in [x_5, x_6] = B$ and $\exists \epsilon > 0$ such that $f(x) \leq f(y)$ where $y \in [x_5, x_6]$ and $x \in (x_5 - \epsilon, x_6 + \epsilon)$. In the procedures stated in the following sections, we will be satisfied if one or more points belonging to $[x_5, x_6]$ may be termed as local best representative point.

Note that the above example has been stated for density functions in $\Re$. The stated concepts are also valid for $\Re^2, \Re^3, \cdots$.

Note also that the basic idea behind local best representative points is somewhat similar to the notion of local *mode*, but there exists some difference. For uniformly distributed data, a *mode* does not exist but we can choose a representative point [as in 3) of the above example]. These concepts are formally defined below.

Definition 1 (D.1): For a set α with probability measure Q_α and density function f_α, a point $x_0 \in \alpha$ is said to be a *local representative point* if there exists an open set $\mathcal{V}$ containing x_0 such that $f_\alpha(x_0) \geq f_\alpha(x) \qquad \forall\, x \in \mathcal{V}$. The point x_0 is a *local best representative point* if $f_\alpha(x_0)$ is strictly greater than $f_\alpha(x)$ for all x.

Definition 2 (D.2): For a set α with probability measure Q_α and density f_α, a point $x_0 \in \alpha$ is said to be a *global best representative point* of α if $f_\alpha(x_0) > f_\alpha(x) \qquad \forall\, x \in \alpha$ and $x \neq x_0$.

The definitions of global and local best representative point in the finite number data points are given below.

Definition 3 (D.3): Let $S = \{x_1, x_2, \cdots, x_n\}$ be the pattern set of n points. A point x_0 is said to be a *global best representative point* of S if

$$\#(V_{x_0,\theta} \cap S) > \#(V_{x,\theta} \cap S)$$
$$\forall\, x \in S, x \neq x_0$$

where $V_{a,b}$ is an open set with center at a and radius b, and $\#$ denotes the number of points.

Note that Definition 3 intuitively reflects the basic idea of Definition 2. Note also that no mention has been made about the value of θ. It may also be observed in the above definition that x_0 may or may not belong to S.

The definition of *local best representative point* in the sample, which is to be followed in this correspondence, differs slightly from the earlier definition.

Definition 4 (D.4): Let $S = \{x_1, x_2, \cdots, x_n\}$ be the set of n random points. A point x_0 is said to be a *local best representative point* if

$$\#(V_{x_0,\theta} \cap S) > \#(V_{x,\theta} \cap S)$$
$$\forall\, x, x_0 \in S; x \neq x_0.$$

We have already noted that there is no guideline about the value of θ. The radius of the open disc, namely θ, is an important impediment for detecting local best representative points. If the radius is very large, then all the points in the data set may lie within the disc and, in this case, it may not be possible to decide which one is the local best representative point. If the radius is very small, then the open disc may not contain enough number of data points to be amenable to make any decision [13]. In this connection, a measure of finding the radius is explained below.

As a consequence of the above discussion, the radius θ should depend on the interpoint distances. We assume that θ is a function of average of the square of the edge weights of MST of the pattern set, namely S. In other words, θ is a function of ℓ_n/n, where ℓ_n is the sum of the edge weights of MST of S (edge weight being the Euclidean interpoint distance). In our experiments, we take the initial value of θ to be equal to $h_n = \sqrt{\ell_n/n}$. It may be noted here that h_n has been shown to be useful in set estimation too [12]. We state below a result involving h_n and set estimation.

Let $\alpha \subseteq \Re^2$ be path connected, compact set with $cl(Int(\alpha)) = \alpha$ and $\mu(\delta\alpha) = 0$ where $Int(A)$ represents interior of A, $cl(A)$ represents closure of A, and δA represents the boundary of A, i.e., $\delta A = cl(A) \cap cl(A^c)$, and μ represents the Lebesgue measure in $\Re^2$.

Let $x_1, x_2, \cdots, x_n$ be independent random vectors in $\Re^2$ identically distributed with probability measure Q_α, i.e., $Q_\alpha(A) > 0$ if A is open and $A \cap \alpha \neq \emptyset$. Let $h_n = \sqrt{\ell_n/n}$, where ℓ_n is the sum of edge weights of minimal spanning tree (MST) [7] of $x_1, x_2, \cdots, x_n$, where the Euclidean interpoint distance is taken as the edge weight.

Let $\alpha_n^* = \cup_{i=1}^n \{x : \|x_i - x\| \leq h_n\}$. Then α_n^* is a consistent estimator of α. In other words, $E_\alpha[\mu(\alpha_n^* \Delta \alpha)] \to 0$ as $n \to \infty$, where μ is the Lebesgue measure in $\Re^2$, E represents expectation, and Δ represents the symmetric difference between sets [12].

III. Proposed Algorithms

At first, we assume that n_0 is not supplied. In such a situation, the algorithm for obtaining *local best representative points* from $S = \{x_1, x_2, \cdots, x_n\} \subseteq \Re^2$ is described below.

Algorithm AL-1:

Step 1: Find the radius h_n as discussed before.

Step 2: Compute the *density* (the number of points) for each datum x as the number of other data units within an open disc of radius h_n with x as center. Let m_i denote the density of the point

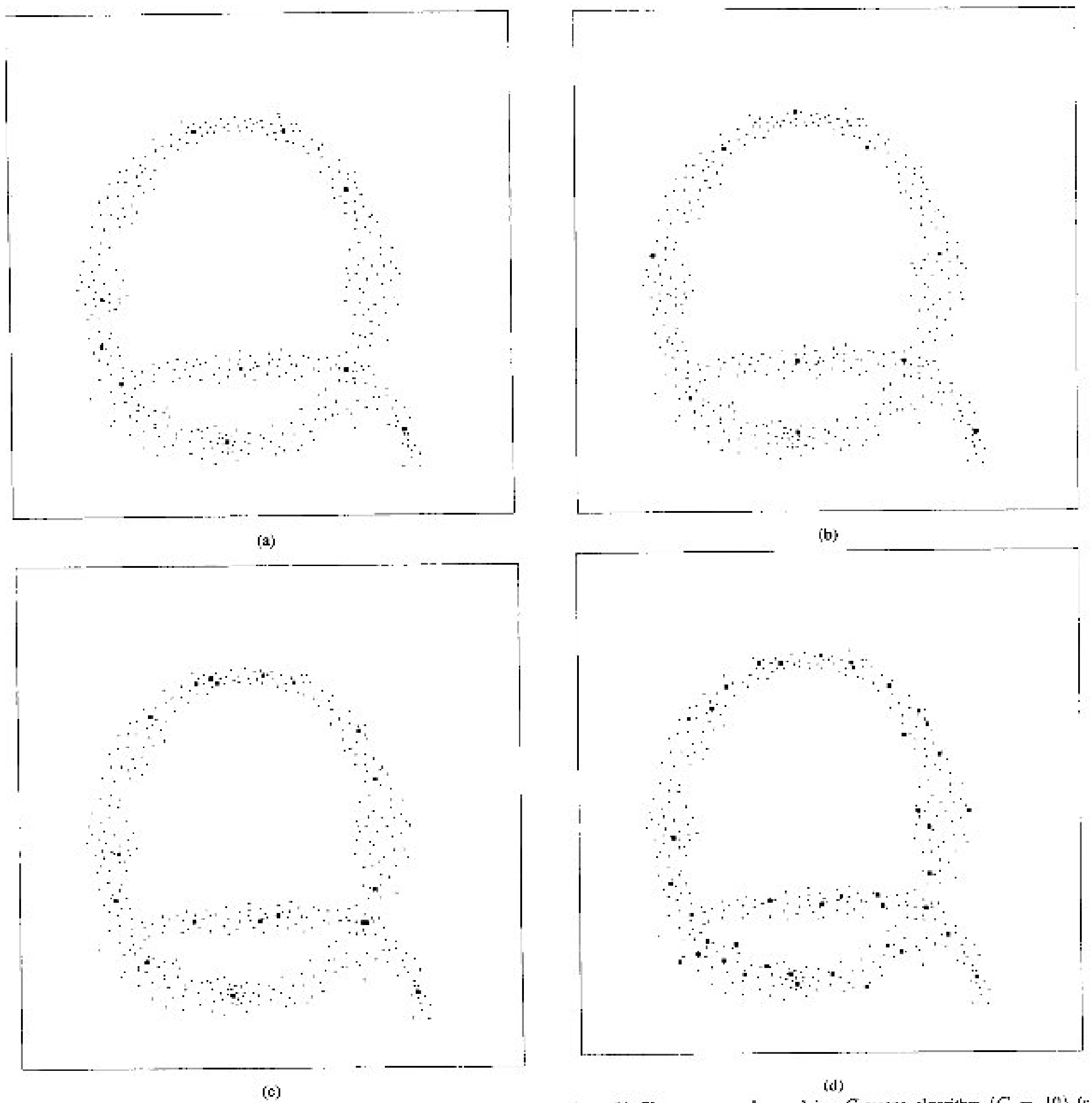


Fig. 6. Q-shaped synthetic data of size 412. (a) Local best representative points. (b) Cluster centers by applying C-means algorithm ($C = 10$) (c) Almost 5% representative points. (d) Almost 10% representative points.

$x_i; i = 1, n$. In other words, let $A_i = \{y : ||x_i - y|| \leq h_n, y \in S\}, i = 1, 2, \cdots, n$, and $m_i = \#A_i, i = 1, 2, \cdots, n$ ($\#A$ means the number of points of the set A).

Step 3: Rearrange $m_1, m_2, \cdots, m_n$ in increasing order. Let the rearrangement be $m_1^*, m_2^*, \cdots m_n^*$. Let $p_j, j = 1, n$ represent the corresponding cumulative sums of $m_1^*, m_2^*, \cdots, m_n^*$; i.e., $p_j = \Sigma_{i=1}^{j} m_i^*, j = 1, 2, \cdots, n$.

Step 4: Compute $M = [(w/100) \times n]$, where $[a]$ means integral part of "a," i.e., the largest integer $\leq a$. Find the value of i for which p_i is nearest to M. Choose $k = m_i^*$ for that i. If

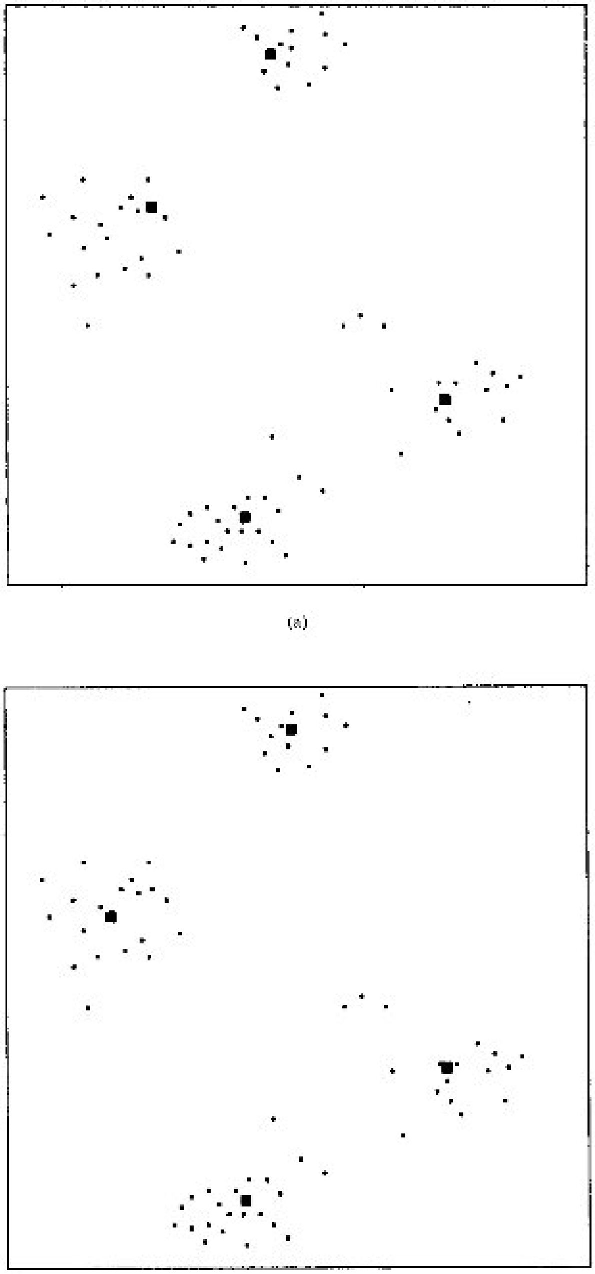
$$p_i < M < p_{i+1}$$

and $M - p_i = p_{i+1} - M$, then choose $k = m_{i+1}^*$. The value of w is guided by the possible use of the representative points. We have used them as seed points for clustering and found that $w = 85$ gives consistently good results.

Step 5: Find the set $S_1 \subseteq S$ such that every point in S_1 has density at least equal to k: i.e., $S_1 = \{x_i : m_i \geq k, x_i \in S\} \subseteq S$.

Step 6: Choose any point of S_1 as the first *initial* local best representative point of S. For convenience, we choose that point of S_1 whose suffix is minimum as the first initial local best representative point.

Step 7: Choose subsequent *initial* local best representative points from S_1 subject to the stipulation that each new local best representative point is at least at a distance $2h_n$ from all other previously chosen initial local best representative points.



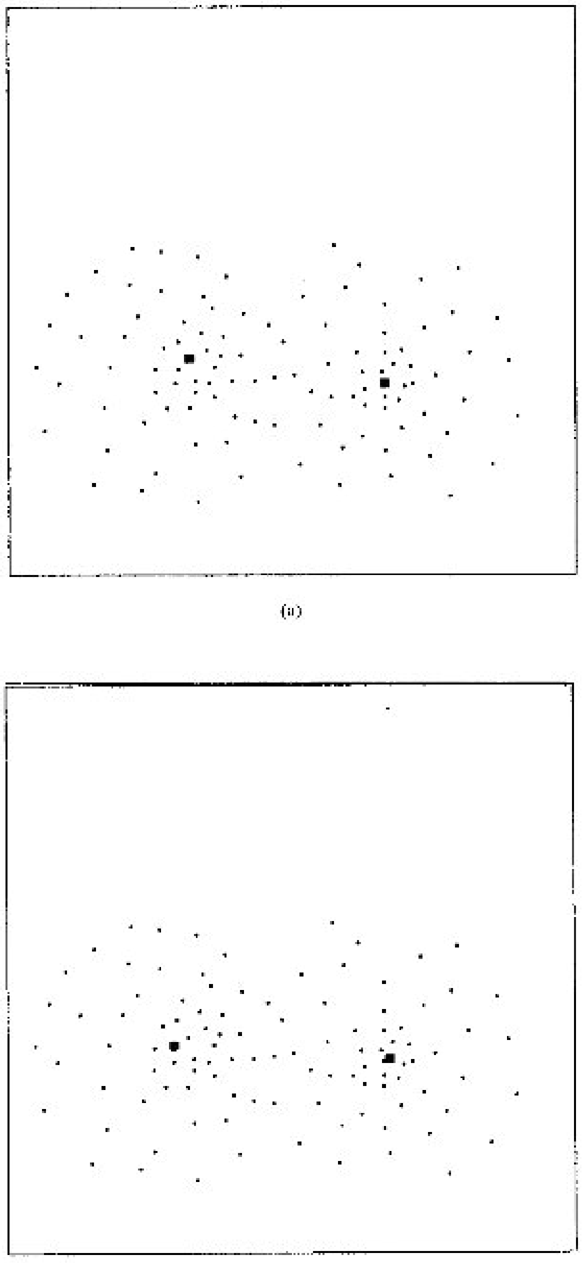
(a)

(b)

Fig. 7. Ruspini's data of size 75. (a) 4 local best representative points. (b) 4 cluster centers by applying C-means algorithm ($C = 4$).

Continue choosing initial local best representative points until all remaining data units of S_1 are exhausted. Let V be the set of initial local best representative points of S. Let $t = \#V$. Let $V = \{y_i, i = 1, 2, \cdots, t\}$.

Step 8: Find the sets B_i which are the collection of those points, from S_1, whose distances from the ith initial local best representative point are less than or equal to $2h_n$, for $i = 1, 2, \cdots, t$; i.e., $B_i = \{z : ||z - y_i| \leq 2h_n, z \in S_1\}, i = 1, 2, \cdots t$.



(a)

(b)

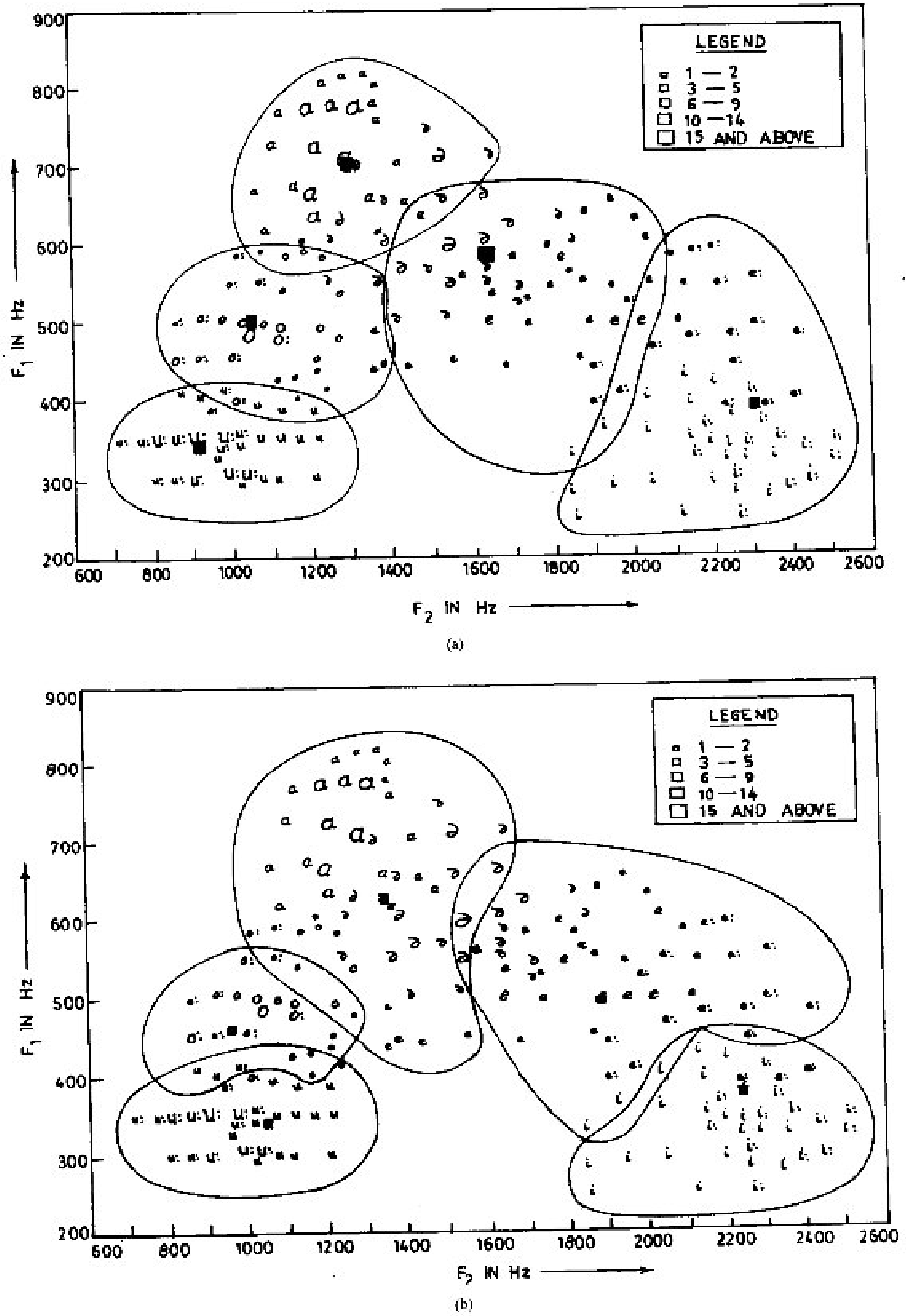
Fig. 8. Overlapping Gaussian cluster data of size 107. (a) 2 local best representative points. (b) 2 cluster centers by applying C-means algorithm ($C = 2$).

Step 9: Find the *density* for each data unit of the sets $B_i, i = 1, 2, \cdots, t$. Find the point x_i^* of B_i, for every $i = 1, 2, \cdots t$ whose *density* is maximum in B_i.

Consider the subset

$$V^* = \{x_1^*, x_2^*, \cdots, x_t^*\} \subseteq S.$$

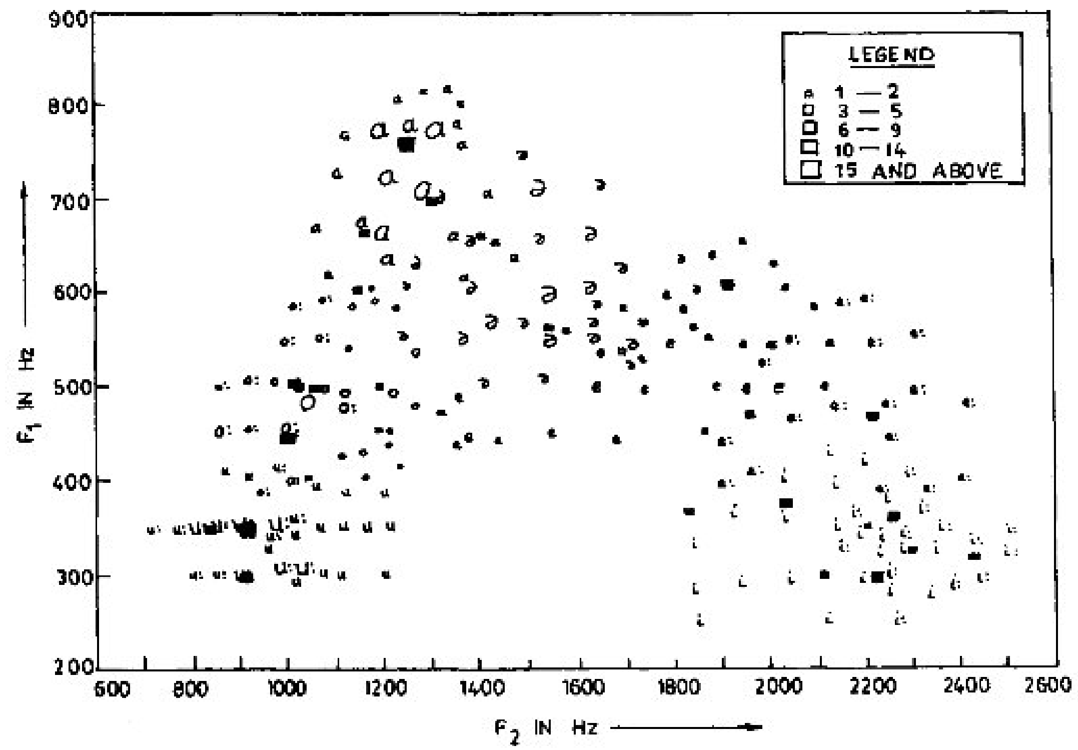
V^* is the output of the algorithm representing the local best representative points.



(a)

(b)

Fig. 9. Vowel sound data set (size of the character represent density according to the legend). (a) 5 local best representative points. (b) 5 cluster centers by applying C-means algorithm ($C = 5$).



(c)

Fig. 9. (c) Almost 5% representative points.

Step 10: Stop.

In Step 4, M is used to discard the data of low density from our consideration. Note that k and S_1 are dependent on M. If M increases (decreases), then the value of k will also increase (decrease) and the number of points in S_1 will decrease (increase).

When n_0 is supplied, Algorithm AL-2 may be used to find the representative points.

Algorithm AL-2:

Step 1: Find h_n and take it as radius θ.

Step 2: Compute the *density* (the number of points) for each data unit as the number of other data units within open disc of radius θ. Let $m_1, m_2, \cdots m_n$ be the number of points corresponding to n centers $x_1, x_2, \cdots, x_n$. That is, let $A_i = \{y : ||x_i - y|| \leq \theta, y \in S\}, i = 1, 2, \cdots, n$ and $m_i = \#A_i, i = 1, 2, \cdots, n$.

Step 3: Apply Step 3 to Step 9 of AL-1.

Step 4: If $t < n_0$, then go to Step 5. If $t > n_0$, go to Step 6. Else, go to Step 7.

Step 5: Decrease θ in a predefined manner and go to Step 2.

Step 6: Increase θ in a predefined manner and go to Step 2.

Step 7: Stop.

The flowchart of the above algorithm is in Fig. 2.

The experimental results for two-dimensional data sets are described in the next section.

IV. Experimental Results for 2-D on Synthetic Data

For our convenience, the results of AL-1 are called seed points. The algorithms AL-1 and AL-2 described in this correspondence have been implemented on data sets of different size and shape.

Fig. 3(a) shows a triangular distribution data of size 200. Here $h_n = 0.4576$. By algorithm AL-1, only one seed point has been found [marked by ■ in Fig. 3(a)]. After applying C-means [4] algorithm (here $C = 1$), we see that the cluster center in Fig. 3(b) is very near to the intitial seed point. If we start with our initial seed point detected by AL-1, the C-means algorithm converges after only one iteration. So, our algorithm chooses a good seed point.

Next, we consider the problem of choosing nearly 5 and 10% representative points. AL-2 was used to detect these points. The results shown in Figs. 3(c) and (d), respectively, have been obtained when θ is reduced from h_n by 10 and 20%, respectively.

The results of the algorithm on other data sets are also demonstrated. Fig. 4(a) shows a data of two clusters of size 322. Here $h_n = 0.5631$. As shown in Fig. 4(a), two representative points are found by using AL-1. Again, note the nearness of cluster centers obtained by C-means ($C = 2$) algorithm [Fig. 4(b)]. Almost 5% and 10% representative points are shown in Fig. 4(c) and (d), respectively.

The algorithms are tested on point patterns of nonconvex shapes as well. Fig. 5(a) shows C-shaped data of size 214 with $h_n = 0.5763$. Fig. 6(a) shows Q-shaped data of size 412 with $h_n = 0.6534$. Note that the seed points detected in both cases are visually acceptable. We used a combination of dot patterns in Figs. 4–6, and obtained good clustering by a method of multiple seed points. The results will be reported in a future correspondence.

Fig. 7(a) shows Ruspini's data [15] of size 75. Fig. 8(a) shows overlapping Gaussian clusters data of size 107. In the former case, four seed points are automatically detected by algorithm AL-1, and in the latter case, two seed points. Note that the number of prominent clusters in these two cases are four and two, respectively.

V. Extension for Three and Higher Dimensions

The extensions of our approach to three or higher dimensions is rather straightforward. In this case, the data should he examined in a hypersphere of radius h_n where

$$h_n = \left(\frac{\ell_n}{n}\right)^{1/2}$$

where ℓ_n is the sum of edge weights of MST of $\{x_1, x_2, \cdots, x_n\} \subseteq \Re^q (q \geq 3)$, and the Euclidean interpoint distance is taken as the edge weight. The rest of the algorithms stated in Section III can be used for higher dimension. An example is stated below for a real life data set where $q = 3$.

The data consist of a set of 871 discrete, phonetically balanced speech units in a vowel vocabulary uttered by three male speakers aged 30–35 years. For vowel sounds of five classes including short and long categories $(\delta, a, i, i:, u, u:, e, e:, o,$ and $o:)$, the first three formant frequencies at the steady state (F_1, F_2,and F_3) are obtained through spectrum analysis. These data, when represented in the three-dimensional coordinates (where each coordinate represents a formant frequency), are subject to representative point detection. Fig. 9(a) shows the feature space of vowels corresponding to F_1 and F_2 when longer and shorter categories are treated separately. Here $h_n = 132.7324$. By applying the AL-1 algorithm, five seed points of this data set are found. Note that they correspond to five major vowels when *long* and *short* versions of a vowel belong to the same category. Using these seed points, the C-means ($C = 5$) algorithm terminates in three iterations. See also the 5% representative points obtained using AL-2 [Fig. 9(c)].

VI. Discussion

The problem of choosing a subset of representative data from a larger set is considered in this correspondence. However, there exists no standard definition of the representative point. As stated before, representative points could be defined so as to minimize the root mean square error in the representation. If S_0 denotes the subset of representative points, then for each $p \in S_0$, let $T(p) \subseteq S$ denote the subset of nearest neighbors of p. The root mean square error may be defined as

$$E = \sum_{p \in S_0} \sum_{q \subset T(p)} d(p, q)$$

where $d(p, q)$ is the Euclidean distance of p from q. The best representatives can be given by the subset S_0 for which E is minimum.

Unfortunately, there exist ${}^nC_{n_0}$ possibilities of choosing S_0. Thus, the algorithm has an order of complexity n^{n_0} which is computationally impractical for any reasonable number of n and n_0. Our method, which is based on local density estimate, is essentially an $O(n^2)$ algorithm. Here, the notion of *good* representatives is identical with the data with high density.

The representative points found by our algorithms can be used in various applications including clustering and K-nearest neighbor classifier. A multiple seed point clustering algorithm is being examined by us, and important results will be communicated in a future correspondence.

Our approach and algorithm have been tested on synthetic as well as real life data, and the results appear to be satisfactory. However, it is interesting to find theoretical properties of the approach. The work is in progress, and the useful results will be reported in a future correspondence.

One of the disadvantages of the algorithms lies in heuristic choice of w. Thus, if the number of representative points n_0 is fixed *a priori*, there is no guarantee of obtaining exactly n_0 points by these algorithms. One has to vary w and see when the number of resulting points is near n_0. An alternative approach [17], [18] has been proposed by us to find exactly n_0 representatives.

Acknowledgment

Interest in the work by Prof. D. Dutta Majumder is acknowledged with thanks. The authors also with to thank S. Chakraborty for his drawings.

References

[1] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, 1962.

[2] T. Cacoullos, "Estimation of a multivariate density," *Ann. Inst. Statist. Math.*, vol. 18, pp. 179–189, 1966.

[3] J. B. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Symp. Math. Statist. Probab.*, vol. 1, Berkeley, CA, 1967, pp. 281–297, AD669871.

[4] E. W. Forgy, "Cluster analysis of multivariate data: Efficiency versus interpretability of classifications," presented at the Biometric Soc. Meet., Riverside, CA, 1965, (Abstract in *Biometrics*, vol. 21, no. 3, p. 768.)

[5] M. M. Astrahan, "Speech analysis by clustering, or the hyperphoneme method," Stanford Artif. Intell. Proj. Memo. AIM-124, AD 709067, Stanford Univ., Stanford, CA, 1970.

[6] G. H. Ball and D. J. Hall, "PROMENADE –An online pattern recognition system," Rep. RADC-TR-67-310, AD822174, Stanford Res. Inst., Menlo Park, CA, 1967.

[7] C. T. Zahn, "Graph theoretic methods for detecting and describing gestalt clusters," *IEEE Trans. Comput.*, pp. 68–86, 1971.

[8] M. R. Anderberg, *Cluster Analysis for Application*. New York: Academic, 1973.

[9] P. A. Devijver and J. Kittler, *Pattern Recognition- A Statistical Approach*. London: Prentice Hall International, 1982.

[10] B. Kleiner and J. A. Hartigan, "Representing points in many dimensions by trees and castles," *J. Amer. Statist. Assoc.*, vol. 76, pp. 260–276, 1981.

[11] R. F. Ling, "A probability theory of cluster analysis," *Amer. Statist. Assoc.*, vol. 68, no. 3, pp. 159–164, 1973.

[12] C. A. Murthy, "On consistent estimation of class in $\Re^2$ in the context of cluster analysis," Ph.D. dissertation, I.S.I., Calcutta, 1989.

[13] D. Chaudhuri, B. B. Chaudhuri, and C. A. Murthy, "A new split-and-merge clustering technique," *Pattern Recognition Lett.*, vol. 13, no. 6, pp. 399–409, 1992.

[14] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[15] H. R. Ruspini, "Numerical methods for fuzzy clustering," *Inform. Sci.*, vol. 2, pp. 319–350, 1970.

[16] D. Dutta Majumder, A. K. Datta, and N. R. Ganguli, ""Some studies on acoustic phonetic features of Telugu vowels," *Acoustica*, vol. 41, no. 2, pp. 55–64, 1978.

[17] B. B. Chaudhuri, "How to choose a representative subset from a set of data in multi-dimensional space," *Pattern Recognition Lett.*, accepted for publication.

[18] B. B. Chaudhuri and D. Chaudhuri, "Detection of representative points from a data set," *Tech. Rep.*, TR/KBCS/2/93, 1993.