MINIMAL SPANNING TREE BASED CLUSTERING TECHNIQUE: RELATIONSHIP WITH BAYES CLASSIFIER

NIRMALYA CHOWDHURY[†*] and C. A. MURTHY[‡]

[†]Department of Electrical Engineering, Ramakrishna Mission Shilpapitha, Belghatia, Calcutta 700 035, India
[‡]Machine Intelligence Unit, Indian Statistical Institute, 203 Barrackopore Trunk Road, Calcutta 700 035, India

(*Received* 27 *June* 1996; *in revised form* 5 *November* 1996)

Abstract—A minimal spanning tree (MST) based clustering technique along with its theoretical formulation is presented in this paper. The proposed technique is compared with Bayes Classifier and it is shown theoretically that the clustering technique, although an unsupervised one, approaches the performance of Bayes Classifier under a condition, as the number of sample points from each class increases. Experimental results with many synthetic data sets in 2-D and 3-D validate the theoretical prediction.

Pattern recognition Clustering Minimal spanning tree Bayes classifier
Triangular distribution Truncated normal distribution Error probability

1. INTRODUCTION

Pattern recognition refers to the classification or description of objects or patterns.[1-10] There are essentially two basic types of techniques in Pattern recognition: Supervised and Unsupervised. Supervised pattern recognition techniques are used for the cases where *a priori* information for each class under consideration is available. Here the problem is to assign labels to every pattern vector in the whole feature space. Bayes classifier is a fundamental classifier in supervised pattern recognition. It provides the minimal misclassification probability among all possible partitions of the whole feature space into the given number of classes. It can be implemented easily provided class conditional densities and *a priori* probabilities of the classes are known. But generally, these are not known. Usually researchers try to estimate *a priori* probabilities and densities from a training sample set of patterns and these training sample patterns are assumed to represent the classes properly.[3] A training sample set of patterns consists of finitely many patterns where a class label is attached to each pattern in that set.

In some applications, only a set of training patterns of unknown classification may be available and the user is supposed to provide a label to each pattern. Clustering techniques are unsupervised techniques which attempt to solve the problem by finding structures in the data set and thus partition it into K clusters where K may be known or unknown.[11] Several clustering techniques are available in literature and they use different objective functions for optimization to result in different partitions of the data set. Thus different clustering techniques, many times, result in different cluster configurations and thus making it imperative to validate clustering techniques/clusters. There are several cluster validation techniques existing in the literature.[5]

A probable way of validation is to compare the clusters obtained with the classes obtained from the Bayes classifier, if it is possible. If such a comparison is possible, then it would not only validate the clusters but also provide an objective justification for the corresponding clustering technique. Usually such a comparison is not attempted, because clustering techniques basically explore the data for possible clusters whereas supervised classification techniques utilize either the training sample set or the density functions to result in "good" classification. In the literature, comparison between supervised and unsupervised techniques has not been attempted, because the intuitions and methodologies for these two types of techniques are different. In this article, we aim to show that a common meeting ground for these two techniques does exist and a comparison is feasible under such a situation. The mathematical details presented in this paper will show how such a comparison between supervised and unsupervised techniques is possible. In fact, in this article a clustering scheme is presented and it has been shown theoretically that the resulting cluster boundaries would provide the Bayes decision boundaries as the number of sample points go to infinity under a "smooth" condition, where the number of features is greater than or equal to 2.

In supervised classification, the number of classes (represented by K) is known, whereas in unsupervised techniques, the number of clusters may not be known. The proposed clustering technique finds the number of clusters automatically (i.e. the number of clusters is not an input to the algorithm). It is shown mathematically that as the number of observations (n) goes to infinity, the number of obtained clusters tend to the number of classes and the boundaries of the clusters tend to the correspond-

* Author to whom correspondence should be addressed.

ing Bayes decision boundaries between the classes. The organization of this article is as follows. Section 2 provides the mathematical preliminaries associated with the article. Section 3 presents the mathematical formulation of the problem. The proposed clustering technique is then depicted in Section 4. Experimental results with several synthetic data sets in 2-D and 3-D are presented in Section 5. Concluding remarks are incorporated in Section 6.

2. MATHEMATICAL PRELIMINARIES

The basic problem attempted in this paper is pattern classification. In order to provide a solution one needs to define the phrase "Pattern Class" properly. The definition of pattern class is given below.

2.1. Pattern class

In most of the real life problems, pattern classes are bounded. Thus the pattern classes considered here are also bounded. A formal definition of pattern class in R^N is given below using topological and measure theoretic concepts.

Definition 1. A set $A \subset R^N$ is said to be a pattern class[12] if

(i) A is path connected and compact,

(ii) $\mathrm{cl}(\mathrm{Int}(A)) = A$, [cl means closure, Int means interior]

(iii) $\mathrm{Int}(A)$ is path connected and

(iv) $\mu(\delta A) = 0$ where $\delta A = A \cap \mathrm{cl}(A^c)$ and μ is the Lebesgue measure on R^N.

The relevance of the properties (i)–(iv) of Definition 1 is provided in reference (12). This definition has been used in several other articles[13–15] too. Let $\mathscr{B} = \{A : A$ satisfies Definition 1$\}$. $\mathscr{B}$ is the collection of all classes in R^N. Any $A \in \mathscr{B}$ is referred to as the pattern class. Note that N is the number of features under consideration and the value of N is taken to be greater than or equal to 2.

2.2. Class conditional density function

Definition 2. Let $A \in \mathscr{B}$ be a pattern class. A function f. $f : R^N \to [0, \infty)$ is said to be a class conditional density function on A if

(i) $\int_C f(x)\,dx > 0\ \forall C$, C open, $C \subseteq A$,

(ii) $\int_A f(x)\,dx = 1$,

(iii) f is continuous on A,

(iv) $f(x) > 0\ \forall x \in A$ and

(v) $f(x) = 0\ \forall x \in A^c$.

Explanation: In the above definition, (ii) is the usual property associated with any density function. The property (i) is necessary since, if the probability of an open set $C \subseteq A$ is zero then the class is nothing but $A \cap C^c$ instead of A. Properties (iii) and (v) in the above definition are usual properties associated with density function. Property (iv) is a stringent property assumed on the density function. Many density functions provide zero densities at the boundary points of the set A. Thus $f(x) > 0$, $\forall x \subset \mathrm{Int}(A)$ is a more appropriate property associated with density function. However in a remark made in Section 3, it has been stated that the results in this article would hold in case of "$f(x) > 0, \forall x \in \mathrm{Int}(A)$" also. For mathematical simplicity, the property (iv) is included in Definition 2.

2.3. Mixture density function

Let $A_1, A_2, \ldots, A_K$ be the K pattern classes (as defined in Definition 1) and $p_1, p_2, \ldots, p_K$ be the corresponding class conditional densities (as defined in Definition 2). Let $P_1, P_2, \ldots, P_K$ represent the *a priori* probabilities of the classes $A_1, A_2, \ldots, A_K$ respectively. Let $0 < P_i < 1$, $\forall i = 1, \ldots, K$. Note that $\sum_{i=1}^{K} P_i = 1$. The mixture density function, represented by $p(x)$ is defined as

$$p(x) = \sum_{i=1}^{K} P_i p_i(x).$$

Note. The above definition is the standard definition of a mixture density function.[4]

2.4. Bayes classifier

Note that $A_i = \{x : p_i(x) > 0\}$, $\forall i = 1, 2, \ldots, K$, where A_i is the ith class.

Let $B_1 = \{x : 0 < p_1(x), \quad P_1 p_1(x) \geq P_i p_i(x), \forall i = 2, 3, \ldots, K\}$.

Let $B_i = \{x \in (B_1 \cup B_2 \cup \cdots \cup B_{i-1})^C : 0 < P_i p_i(x), P_i p_i(x) \geq P_j p_j(x), \forall j \neq i\}$, $i = 2, 3, \ldots, K$.

Let $B_i^0 = \mathrm{Int}(B_i)$, $\forall i = 1, 2, \ldots, K$.

It is assumed in this article that each B_i and B_i^0 is a connected set. In most of the PR problems where the number of features $N \geq 2$, this assumption holds. The implication of this assumption will be clear in the later sections. Observe that $\cup A_i = \cup B_i$. Here A_i and B_i denote the regions corresponding to the actual ith class and the Bayes ith class respectively. It may also be noted that $B_i^0 \cup B_j^0$ is a disconnected set for all $i \neq j$.

Note that B_i is the acceptance region for the class i obtained using the Bayes classifier. These regions provide the minimal Bayes error probability e which is given by

$$e = \sum_{i=1}^{K} P_i \int_{B_i^c} p_i(x).$$

2.5. Bayes decision boundary

Generally Bayes decision boundary between the classes i and j $(i \neq j)$ is given by β_{ij}, where

$$\beta_{ij} = \{x : P_i p_i(x) - P_j p_j(x) \geq P_l p_l(x), \ \forall l \neq i, j\}.$$

For the sake of convenience, the set β_{ij} is divided here into two parts β_{ij1} and β_{ij2}, where

$$\beta_{ij1} = \{x : 0 < P_i p_i(x) = P_j p_j(x) > P_l p_l(x), \ \forall l \neq i, j\}$$

and

$$\beta_{ij2} = \{x : 0 = P_i p_i(x) = P_j p_j(x) > P_l p_l(x), \ \forall l \neq i, j\}.$$

Note that $\beta_{ij2} = \{x : P_l p_l(x) = 0, \forall l = 1, 2, \ldots, K\}$. Since $P_l > 0$, $\forall l = 1, 2, \ldots, K$ and $p_l(x) \geq 0$, $\forall x$, and $\forall l = 1, 2, \ldots, K$. In fact $\beta_{ij2} = \{x : p(x) = 0\}$. We shall be considering β_{ij1} as the Bayes decision boundary between the classes i and j in this article. The set β_{ij2} has been handled separately in this article.

2.6. *Random vectors*

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random vectors following the density function $p(x)$ defined above. In other words, there is a set Ω such that

$$X_i : \Omega \to R^N, \forall i = 1, 2, \ldots, n, \ldots$$

Thus for any n, $\{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\}$ is the given data set of n observations, where $\omega \in \Omega$. In other words, a given data set can be represented by $S_n(\omega) = \{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\}$. Note that S_n is a random set in the sense that for a given value of n, one can get several $S_n(\omega)$s depending on the values the random vectors take. Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed random vectors such that

$$Y_i = \begin{cases} 1 & \text{if } X_i \text{ is misclassified according to Bayes rule,} \\ 0 & \text{otherwise.} \end{cases}$$

Then the average number of misclassified points in a given sample of size n is $\frac{1}{n}\sum_{i=1}^{n} Y_i$. From strong law of large numbers we get

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \to e \quad \text{a.c. as } n \to \infty.$$

If $(\gamma_1, \gamma_2, \ldots, \gamma_K)$ represent any other partition of the whole set, such that γ_i is the region corresponding to class i, then the error probability corresponding to this partition is

$$\sum_{i=1}^{K} P_i \int_{\gamma_i^c} p_i(x)\, dx$$

and it is always greater than or equal to e for all such partitions $(\gamma_1, \gamma_2, \ldots, \gamma_K)$.

Let a clustering on $S_n(\omega)$ has provided $K_n(\omega)$ clusters (K_n may be greater than K, equal to K or less than K). Let the $K_n(\omega)$ clusters be represented by $S_{nj}(\omega)$, $j = 1, 2, \ldots, K_n(\omega)$.

Let $C_{nij}(\omega) = S_{nj}(\omega) \cap B_i$, $\forall i = 1, \ldots, K$ and $\forall j = 1, 2, \ldots, K_n(\omega)$. The cluster $S_{nj}(\omega)$ is assumed to correspond to the class i if $\#C_{nij}(\omega) > \#C_{ni_0 j}(\omega)$, $\forall i_0 \neq i$ (where # denotes the number of points). In case of a tie, the decision is taken arbitrarily.

3. FORMULATION OF THE PROBLEM

We would like to find a clustering technique which provides the Bayes classifier. In other words, the partition of the data set should correspond to Bayes classes. Since the number of clusters is unknown, a clustering technique may not result in K clusters. The number of clusters obtained may be greater than K or less than K. As the number of points $n \to \infty$, the number of clusters obtained need to go towards K as well as the obtained clusters should correspond to the Bayes classes. Note that, no prior knowledge on the density of the classes is available. An approximation of any density function may be taken to be the average number of sample points in a disk. This approximation may be obtained if the radius of the disk is "suitably" chosen. Observe that, from the sample, we can at most approximate the mixture density function, but not the class conditional density functions. Thus the approximation of the mixture density function need to be used to get the Bayes classes. A meaning for the "mixture densities providing the Bayes classes" is that the "valley points" in the mixture density function should correspond to Bayes decision boundaries between the classes. In fact, there needs to be a relationship between the mixture density function and the Bayes classifier if a clustering method should result in Bayes classifier.

A way of defining valley points in a density function, an assumption relating to valley points are stated in Section 3.1.

3.1. *Valley points*

A point $x_0 \in R^N$ is said to be a valley point of $p(x)$ if $\exists r > 0$, and $y_1, y_2 \in R^N$ such that

(i) $\exists l_0$ $(0 < l_0 < 1)$ such that $x_0 = l_0 y_1 + (1 - l_0) y_2$;

(ii) $\forall l \neq l_0$, $0 < l < 1$, $\forall x = l y_1 + (1 - l) y_2$ and $\forall r_1 < r$, either

$$0 < \int_{y \in v(x_0, r_1)} p(y)\, dy < \int_{y \in v(x, r_1)} p(y)\, dy \qquad \text{(a)}$$

or

$$\int_{y \in v(x_0, r_1)} p(y)\, dy = 0, \qquad \text{(b)}$$

where $v(x, r)$ is disk of radius r and center at x.

Let V_1 be the set of all $x_0 \in R^N$ satisfying (a) and V_2 be the set of all $x_0 \in R^N$ satisfying (b). Let $V = V_1 \cup V_2$, then V is said to be the set of all valley points of $p(x)$ in R^N.

Remarks. (1) The general idea regarding valley point $x_0 \subset R$ of a function $\alpha(x)$ is that $\alpha(x)$ takes higher values than $\alpha(x_0)$ in a neighbourhood of x_0. Note that neighbourhood of a point in R is an interval. An extension of the concept "interval" to R^2 is the collection of points on a line segment. The set V_1 is constructed using the above mentioned generalization of the "interval". The condition (a) in the definition of V_1 reflects the concept of "valley points" with the help of a line segment. Note that, if no such line segment exists for a point $x \in R^N$, then it is not a valley point intuitively too. Observe that the density of a point in V_1 is greater than zero.

(2) Note that there may exist many "valley points" for which the density is zero. These points are included in V_2

[refer to condition (b)]. In fact, no point in V_2 can belong to any Bayes class.

(3) Note that the constraint (a), stated above, is satisfied by a density function $f \Leftrightarrow f(x_0) < f(x)$, since f is continuous.

The main assumption regarding the valley points is stated below.

Assumption (A1). Let $\delta B_i = \overline{B_i} \cap \overline{B_i^C}$ and $U_{ij} = \delta B_i \cap \delta B_j$ for all $i \neq j$. We assume that

$$\bigcup_{i \neq j} U_{ij} = V_1.$$

Note. The above assumption states that the Bayes boundary between any two classes must belong to the valley region as defined above. This assumption does not contradict the usual real life situations where we find the number of representative samples from both classes to be comparatively less at the boundary between the two classes than that of any other region of any individual class itself. If we do not have valley regions at the boundary between any two classes of a given data set, the proposed technique will not be able to find those classes. Note also that if the boundary region corresponds to high density regions, a clustering technique generally cannot divide the regions in two clusters, since clustering techniques usually attempt to find the high density regions of a given sample and mark each such region as a core of a cluster. This assumption is termed as smooth assumption since it assumes smooth transition of the mixture density function from one class to another class in the feature space.

From now onwards, throughout this article we are going to assume that the Assumption (A1) stated above will be satisfied by the mixture density function. The following results may then be derived under the assumption.

Result 1. $x \in V_2 \Rightarrow x \notin \cup_{i=1}^{K} B_i$.

Proof.
$x \in V_2 \Rightarrow p(x) = 0 \Rightarrow x \notin \cup A_i \Rightarrow x \notin \cup_{i=1}^{K} B_i$.

Result 2. If $x \in V_1$, then x is a point on the Bayes decision boundary between two classes.

Proof. $x \in V_1 \Rightarrow \exists i,j$ such that $x \in \cup_{ij}$
$\Rightarrow x \in \delta B_i \cap \delta B_j \Rightarrow x \in \delta B_i$ and $x \in \delta B_j$
$\Rightarrow P_i p_i(x) - P_j p_j(x) \geq P_l p_l(x), \forall l \neq i,j$
$\Rightarrow x$ is a point on the Bayes decision boundary between ith and jth classes.

Note that the converse is also true, i.e. if x is a point on the Bayes decision boundary between two classes, then $x \in V_1$.

Result 3. If $x \in (V_1 \cup V_2)^C$, then x belongs to exactly one particular Bayes class and x does not belong to boundary of any two classes.

Proof. If $x \notin V_1 \cup V_2 \Rightarrow x \in \cup_{i=1}^{K} B_i$, $x \notin V_1$
$\Rightarrow \exists i$ such that $p_i(x) > 0$ and there cannot exist i_1 and i_2 such that $P_{i1} p_{i1}(x) = P_{i2} p_{i2}(x) > P_l p_l(x)$, $\forall l \neq i_1, i_2$ (from Result 2)
$\Rightarrow$ There exist exactly one i such that $\text{Max}_j P_j p_j(x) = P_i p_i(x)$, i.e. there exists exactly one i such that x belongs to the ith class and x cannot belong to the boundary of any two classes.

Result 4. If $\cup_{ij} = \phi$, then $B_i \cup B_j$ is a disconnected set.

Proof. We need to show that $B_i \cup B_j$ is a disconnected set. Note that B_i and B_j are connected sets. Thus it needs to be proved that either $\overline{B_i} \cap B_j = \phi$ or $B_i \cap \overline{B_j} = \phi$. We shall show below that $\overline{B_i} \cap B_j = \phi$.

Suppose $\overline{B_i} \cap B_j \neq \phi$. Let $i > j$. Let $x \in \overline{B_i} \cap B_j$. Then $0 < P_i p_i(x) - P_j p_j(x) \geq P_l p_l(x)$, $\forall l \neq i,j$. That is, x belongs to the decision boundary between classes i and j.

This implies that $x \in \delta B_i \cap \delta B_j$, which implies that $x \in \cup_{ij}$. This is a contradiction. Thus $\overline{B_i} \cap B_j = \phi$.

Observation. From the above results on valley points, it is clear that, if a method finds $V_1 \cup V_2$, then the Bayes classes may be automatically obtained, since $(V_1 \cup V_2)^C = \cup_{i=1}^{K} B_i^0$ and $B_i^0 \cup B_j^0$ is a disconnected set for all $i \neq j$. The statement of the problem from the above results is stated below.

Statement of the problem. Here, initially, a set of valley points $V_n(\omega)$ of the given data set $S_n(\omega) = \{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\}$ need to be defined and found. It needs to be shown that $V_n(\omega)$ tends to $V_1 \cup V_2$ as $n \to \infty$. The clustering technique needs to find either $V_n(\omega)$ or $V_n^C(\omega)$. The number of disconnected sets in $V_n^C(\omega)$ provide the number of classes as $n \to \infty$ and every connected set correspond to a Bayes class as $n \to \infty$.

4. THE PROPOSED CLUSTERING TECHNIQUE AND ITS CONVERGENCE

It is shown in the previous section that the valley regions in the mixture density function will give rise to Bayes classes under Assumption A1. Thus a technique has been stated below which finds the valley regions in the given data. This method basically utilizes minimal spanning tree of the data set where the edge weight is taken to be the Euclidian distance between the corresponding nodes.

The clustering technique presented in this paper is based on finding the valley regions of the feature space. Valley regions are those regions where the density (the number of points within a given area in 2-D or within a given volume in 3-D) of data points is the lowest compared to that of its neighboring regions. Here the density of any point x is assumed to be the number of data points present in an open disc of radius r around x. Minimal Spanning Tree (MST) of the data points is used to calculate the value of r and Euclidean interpoint distance is taken as the edge weight of the MST. The square-root

of the average edge weight is taken to be equal to r. If we represent the sum of the edge weight of minimal spanning tree of $S_n(\omega)$ by l_n then it can be noted that $l_n \to \infty$ a.e. as $n \to \infty$ and $(l_n/n) \to 0$ in probability.[12] But note that the property "$l_n \to \infty$ a.e. as $n \to \infty$" does not hold good for $N = 1$. Because in such a case, the maximum possible value of l_n would be equal to the length of the class interval. Hence the clustering technique presented in this paper is valid for $N \geq 2$.

In this work we have taken r to be equal to $h_n = (l_n/n)^{1/N}$ where N denotes the dimensionality of the data set. Note that h_n is a function of interpoint distances in $S_n(\omega)$ as well as the number of points n. It may also be noted that $(l_n/n)^{1/N}$ has been considered in other works[16,17] too.

This h_n is used for finding the radius of the disk and consequently the valley regions of the data set. In fact, the proposed technique finds the valley points in the data set by aping the above definition of valley points of a density function by using MST. We shall be considering squares around each point instead of disks. This minor alternation is only for the purpose of ease in implementation. Theoretically, there is no difference between considering squares and considering rectangles.

Let $S_n(\omega) = \{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\}$, $\omega \in \Omega$. That is, the given set of sample points is represented by $S_n(\omega)$. We shall sometimes refer it as S_n for the sake of simplicity. Let for a point $\underline{x} = (x_1, x_2, \ldots, x_N) \subset R^N$ and for a positive real number r,

$$v(\underline{x}, r) = \{(y_1, y_2, \ldots, y_N)' : \|x_i \quad y_i\| \leq r\}.$$

$v(\underline{x}, r)$ is nothing but an N-dimensional square with center at x and side length $2r$. Let $D(\underline{x}, r)$=#$(S_n \cap v(x, r)$. That is, $D(\underline{x}, r)$ denotes the number of points common to $v(\underline{x}, r)$ and S_n. Thus $D(\underline{x}, h_n)$ denotes the number of points common to $v(\underline{x}, h_n)$ and S_n. It has been shown in reference (18) that

$$\frac{D(x, h_n)}{nh_n^N 2^N} \to p(\underline{x})$$

under certain mild conditions. We are going to use this result repeatedly in this section. Note that the conditions under which the above statement is valid have not been stated in this article explicitly in the formulation of the density function (Section 2), because these conditions are not directly related to the main philosophy of the paper. It is also to be noted that most of the density functions satisfy those conditions.

$D(x, h_n)$ is used below in the definition of valley points of a data set S_n. The definition is given here.

Definition 3. Let $D(x, r)$ be the number of points in the data set which fall in a square of side length $2r$ with center at x. Then a point $x_0 \in R^N$ is said to be a valley point of the data set S_n if $\exists y_1, y_2 \in R^N$ such that

(i) $\exists l_0$, $(0 < l_0 < 1)$ such that $x_0 = l_0 y_1 + (1 - l_0) y_2$;

(ii) $\forall l \neq l_0$, $0 < l < 1$ and $\forall x = ly_1 + (1 - l)y_2$ such that either $0 < D(x_0, h_n) < D(x, h_n)$ or $D(x_0, h_n) = 0$.

Let V_n be the set of all such valley points for a given S_n. It will be shown below that the proposed method would give rise to the Bayes classifier as the number of sample points goes to infinity.

We have $V^C = \cup_{i=1}^{K} B_i^0$. Note that B_i^0 is a connected set for all $i = 1, 2, \ldots, K$ and $B_i^0 \cup B_j^0$ is disconnected for all $i \neq j$ and hence the connected components of V^C are $B_1^0, B_2^0, \ldots, B_K^0$. We need to show that $V_n \to V_1 \cup V_2$.

Proof. (i) Let $x \in R^N$ be such that $p(x) = 0$.

$$\exists r_x > 0 \text{ such that } \int_{y \in v(x, r_x)} p(y)\, dy = 0, \quad \forall r < r_x.$$

Which implies that $x \subset V_2$. Now

$$\int_{y \in v(x, r_x)} p(y)\, dy = 0, \quad \forall r < r_x.$$

Which implies that for sufficiently large n, $D(x, h_n) = 0$, since $h_n \to 0$ a.e. $\Rightarrow x \in V_n$ for sufficiently large n.

Thus if $p(x) = 0$ then $x \in V_n$ for sufficiently large $n \Leftrightarrow x \in V_2$.

(ii) Let $x_0 \in R^N$ be such that $p(x_0) > 0$. Let $x_0 \in V_n$ for sufficiently large n, then there exists $x_1, x_2 \in R^N$, $0 < l_0 < l$ and $r_x > 0$ such that

(a) $x = l_0 x_1 + (1 - l_0) x_2$ and

(b) $\forall l \neq l_0$, $0 < l \leq 1$ and $\forall x_0 = lx_1 + (1 - l)x_2$.

This implies that $D(x_0, h_n) < D(x, h_n)$ for sufficiently large n

$$\Rightarrow \frac{D(x_0, h_n)}{nh_n^N 2^N} < \frac{D(x, h_n)}{nh_n^N 2^N} \quad \text{for sufficiently large } n.$$

$$\Rightarrow p(x_0) < p(x) \quad \text{[from reference (18)]}$$

$$\to x_0 \in V_1.$$

(iii) Let $x_0 \in V_1$, then $\exists x_1, x_2 \in R^N$, $0 < l_0 < l$ and $\exists r_x > 0$ such that

(a) $x = l_0 x_1 + (1 - l_0) x_2$ and

(b) $\forall l \neq l_0$, $0 \leq l \leq 1$ and $\forall x_0 = lx_1 + (1 - l)x_2$

$$\Rightarrow p(x_0) < p(x)$$

$$\Rightarrow \frac{D(x_0, h_n)}{nh_n^N 2^N} < \frac{D(x, h_n)}{nh_n^N 2^N} \quad \text{for sufficiently large } n$$

[from reference (18)]

$$\Rightarrow D(x_0, h_n) < D(x, h_n) \text{ for sufficiently large } n$$

$$\Rightarrow x \in V_n \text{ for sufficiently large } n.$$

Thus if $p(x_0) > 0$ then $x_0 \in V_1 \Leftrightarrow x_0 \in V_n$ for sufficiently large n. Combining (i)–(iii), it has been shown above that $V_n \to V_1 \cup V_2$ as $n \to \infty$. Note that similar results can be proved if property (iv) of Definition 2 is modified in the following way:

$$f(x) > 0 \forall x \in \text{Int}(A).$$

5. IMPLEMENTATION DETAILS AND RESULTS

It has been already stated that the clustering technique presented in this paper is based on finding the valley regions of the feature space. Valley regions are those

regions where the density [the number of points within a given area in 2-D (within a given volume in 3-D)] of data points is the lowest compared to that of its neighboring regions. A possible of way of computing the local densities of data points is to employ open disk (in case of R^2) or sphere of radius r (in case of R^3). The number of data points that fall within such a disk or sphere can be assumed to be the density of the region occupied by that disk or sphere. Another way of computing the local densities of data points is to use squares (cubes in 3-D) of side l. For simplicity of computation, in all the experiments for this work, we have considered squares (cubes in 3-D) of side h_n throughout the given feature space, where minimal spanning tree (MST) of the data points is used to decide the value of h_n. The density associated with any such square is taken to be the number of data points falling within that square. A square (cube in 3-D) may have a maximum number of eight squares (26 cubes in 3-D) in its neighborhood.

A square T (to be termed as pixel T) is assumed to be in valley region if its density is less than the density of its neighboring pixels T_{i1}, T_{i2} for at least one i, where i varies from 1 to 4. Graphically, $T_{11}, T_{12}, T_{21}, T_{22}, T_{31}, T_{32}, T_{41}, T_{42}$ and T are shown below (Fig. 1).

In other words, only four lines are considered for 2-D case. Note that for 3 D case, where cubes are taken and then and there are 26 neighbors for each cube, the number of lines under consideration would be 13. For the general N-dimensional case, the number of lines under consideration would be $[(3^N - 1)/2]$. The number of lines to be considered would increase exponentially with N.

Given a data set having n points, initially the proposed method finds the squares that belong to valley regions. Then those squares and the points that are associated with them are removed from the process. Now we have some disjoint sets of squares. The number of such disconnected sets is taken to be equal to the number of clusters present in the data set. The data points that are associated with any square of a particular set of squares are assumed to belong to that particular cluster. Lastly there may exist valley points which would not go to any cluster. For these points one may follow any one of the following strategies:

(a) The points are taken to be not belonging to any cluster.

(b) These points are put in some clusters according to some criterion such as nearest neighbor classifier rule.[19]

Since conventional clustering techniques assign each data point to any one of the possible clusters, hence we have followed the later strategy and assigned the data points that belong to the valley regions to any one of the existing clusters on the basis of nearest neighbor classifier rule. The efficiency of the process is judged by the number of misclassified points.

A way of checking the validity of obtained classification is to calculate the error probability with respect to the obtained classification and showing that this error probability would go to the Bayes error probability $e(\delta)$ [since e is a function of inter class distance δ, so from now onwards it will be denoted as $e(\delta)$]. Note that from strong law of large numbers, one can take the obtained average misclassification to be an approximation of error probability corresponding to that classification. Thus an experimental verification has been made where the observed average misclassification is shown to be going towards the Bayes error probability e by taking some increasing values of n.

The obtained average misclassification for given data of size n is given by

$$e_n(\omega) = \frac{\text{Number of misclassified points}}{n}.$$

Note that, the lower the value of $e_n(\omega)$, the better is the classification.

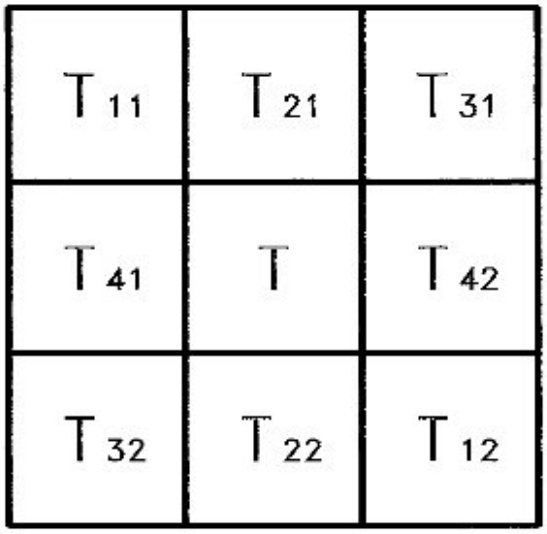

Fig. 1. $T_{11}, T_{12}, T_{21}, T_{22}, T_{31}, T_{32}, T_{41}, T_{42}$ and T.

6. EXPERIMENTAL RESULTS AND ANALYSIS

Several experiments have been carried out on synthetic data sets to judge the validity of the proposed technique. The experiments and their results are described below.

Experiment 1. In this experiment, we have considered several synthetic data sets in R^2. Here the data points are generated from two classes A_1 and A_2 using truncated normal distribution in R^2, where

$$A_1 = [-2, 2] \times [0, 2]$$

and

$$A_{2\delta} = [-2 + \delta, 2 + \delta] \times [0, 2],$$

where $\delta = 3$, 3.5 and 4. Here the x- and y-coordinates of the data points are generated from truncated normal distribution and uniform distribution respectively.

The class conditional density functions for class A_1 and $A_{2\delta}$ are denoted by $p_1(x, y) = f_1(x)f_3(y)$ and

$p_{2\delta}(x,y) = f_{2\delta}(x)f_3(y)$ respectively, where

$$f_1(x) = \begin{cases} \frac{\exp[-(1/2)x^2]}{\sqrt{2\pi}(1-a)} & \text{if } x \in [-2,2], \\ 0 & \text{otherwise,} \end{cases}$$

$$f_{2\delta}(x) = \begin{cases} \frac{\exp[-(1/2)(x-\delta)^2]}{\sqrt{2\pi}(1-a)} & \text{if } x \in [-2+\delta, 2+\delta], \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f_3(y) = \begin{cases} \frac{1}{2} & \text{if } x \in [0,2], \\ 0 & \text{otherwise,} \end{cases}$$

where a is given by

$$\int_{-\infty}^{-2} \frac{1}{\sqrt{2\pi}} \exp[-(1/2)x^2] + \int_{2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[-(1/2)x^2]$$
$$= 2\int_{2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[-(1/2)x^2].$$

The *a priori* probabilities for classes A_1 and A_2 are taken to be 1/2. Thus the Bayes decision boundary between classes A_1 and A_2 is given by

$$p_1(x,y) = p_{2\delta}(x,y) \Rightarrow f_1(x)f_3(y)$$
$$= f_{2\delta}(x)f_3(y) \Rightarrow f_1(x) - f_{2\delta}(x) \Rightarrow x = \frac{\delta}{2}.$$

The Bayes decision boundary between the classes is given by the set $\{(\delta/2, y) : y \in [0,2]\}$. The mixture density function is given by

$$p(x,y) = \frac{p_1(x,y) + p_{2\delta}(x,y)}{2}$$
$$\Rightarrow p(x,y) = \frac{[f_1(x) + f_{2\delta}(x)]f_3(y)}{2}.$$

It can be shown that the function $p_\delta(x,y)$ has a minima at $\{(\delta/2, y) : y \in [0,2]\}$. Since the valley region of $p_\delta(x,y)$ is the same as the Bayes decision boundary between the classes, hence Assumption A1 is satisfied for the data set generated by the method as described above.

Note that the distance between the classes A_1 and $A_{2\delta}$ is taken to be δ, where δ is the difference between the means of the truncated normal distribution for the two classes. Also note that the Bayes error probability $e(\delta)$ associated with any given data set that is generated from A_1 and A_2 is a function of δ, where

$$e(\delta) = \begin{cases} 0 & \text{if } \delta \geq 4, \\ \frac{1}{(1-a)} \int_{\delta/2}^{2} f_1(x)\,dx & \text{if } 0 < \delta < 4. \end{cases}$$

Table 1 presents the results for this experiment. We have considered various values of δ for the experiments and reported the results for $\delta = 4$, 3.5, and 3. For each δ, the different sample sizes considered are 500, 600, 700, 800, 900, 1000, 1100, 1200. The average error $(e_n(\omega))$ found for each sample is as shown in Table 1 for $\delta = 4$, 3.5, and 3. Note that for $\delta > 4$, $e_n(\omega)$ values are all found to be zero which is same as Bayes error probability. The Bayes regions for the two classes for $\delta = 4$ are shown in Fig. 2(a). The bold vertical line in the middle denotes the Bayes decision boundary between the two classes. The regions obtained by the proposed technique for $\delta = 4$ for sample sizes of 500, 900, and 1200 are shown in Fig. 2(b)–(d) for the purpose of visual comparison. It can be seen from Fig. 2(a)–(d) that the regions obtained by the proposed technique is going towards the corresponding Bayes region as the number of data points in the given sample increases. It can also be seen from Table 1 that the average error $e_n(\omega)$ is going towards Bayes error probability $e(\delta)$ as the number of data points in the given sample increases.

Table 1. Results with 2 class problem in 2-D using truncated normal distribution.

δ Value	No. of points	Average Error $e_n(\omega)$	Bayes Error Probability $e(\delta)$
	500	0.2010000	
	600	0.1983500	
	700	0.1992000	
4	800	0.0005625	0
	900	0.0004556	
	1000	0.0004500	
	1100	0.0002636	
	1200	0.0001200	
	500	0.3889796	
	600	0.3305525	
	700	0.2744200	
3.5	800	0.2019656	0.00138354
	900	0.1866012	
	1000	0.1005100	
	1100	0.0609366	
	1200	0.0335400	
	500	0.4999796	
	600	0.4999666	
	700	0.4228598	
3	800	0.3885225	0.04661857
	900	0.3302352	
	1000	0.2532000	
	1100	0.1778225	
	1200	0.0999666	

Experiment 2. In this experiment, we have considered data sets that are generated from three classes A_3, A_4 and A_5 in R^2, where

$$A_3 = [0,2] \times [0,2],$$

$$A_4 = [2,4] \times [0,2]$$

and

$$A_5 = [1,3] \times [2,4].$$

In this case triangular distribution is used. The class conditional density functions for classes A_3, A_4 and A_5 are taken to be $p_1(x,y)$, $p_2(x,y)$ and $p_3(x,y)$ respectively,

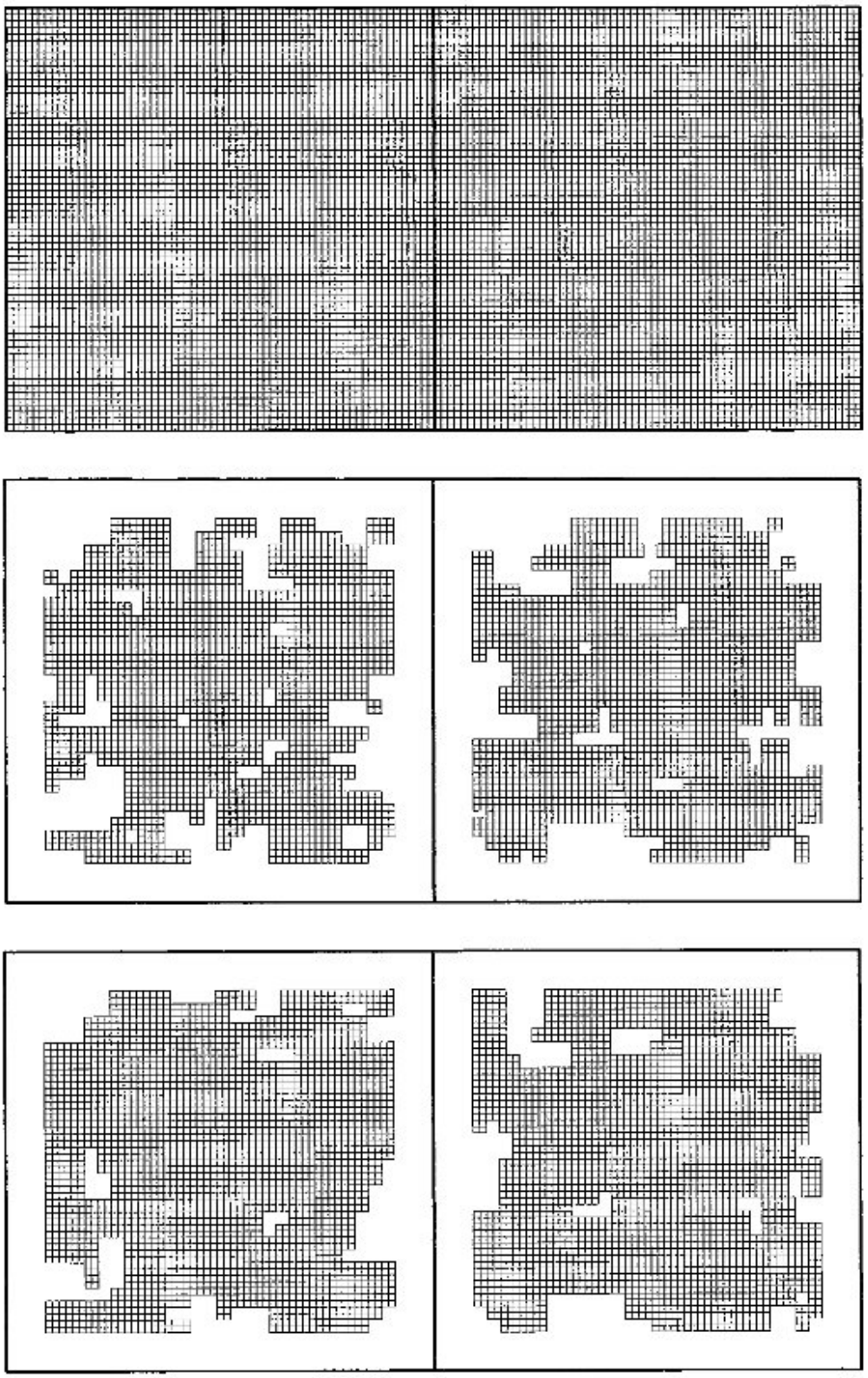

Fig. 2. (a) The Bayes regions for the two classes for δ=4. (b)–(d) The regions obtained by the proposed technique for δ 4 for sample sizes of 500, 900, and 1200.

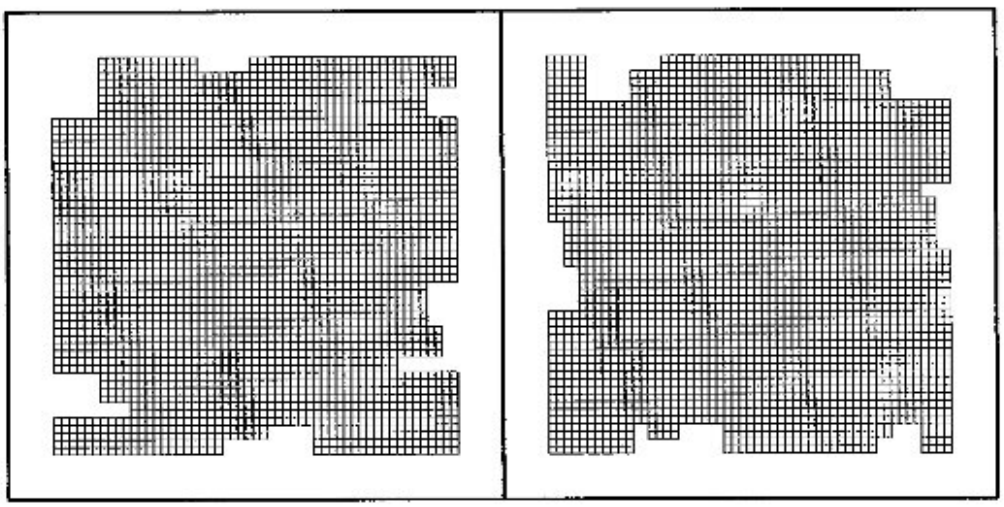

Fig. 2. (Continued).

where

$$p_1(x,y) = f_4(x)f_4(y),$$
$$p_2(x,y) = f_5(x)f_4(y)$$

and

$$p_3(x,y) = f_6(x)f_5(y),$$

where

$$f_4(x) = \begin{cases} x & \text{if } x \in [0,1], \\ 2-x & \text{if } x \in [1,2], \\ 0 & \text{otherwise,} \end{cases}$$

$$f_5(x) = \begin{cases} x-2 & \text{if } x \in [2,3], \\ 4-x & \text{if } x \in [3,4], \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f_6(x) = \begin{cases} x-1 & \text{if } x \in [1,2], \\ 3-x & \text{if } x \in [2,3], \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the three classes A_3, A_4 and A_5 considered for the experiments are taken to be non-overlapping. Here the *a priori* probabilities for the three classes are taken to be equal to $\frac{1}{3}$. Thus Assumption A1 stated above is also valid in this case. Table 2 depicts the results of this experiment. It can be seen from Table 2 that the average error $e_n(\omega)$ is going towards Bayes error probability $e(\delta)$ as the number of data points in the given sample increases.

Table 2. Results with 3 class problem in 2-D using triangular distribution

δ Value	No. of points	Average Error $e_n(\omega)$	Bayes Error Probability $e(\delta)$
	500	0.66533	
	600	0.66150	
	700	0.53480	
0	800	0.46599	0
	900	0.39828	
	1000	0.14000	
	1100	0.03377	
	1200	0.00800	

Experiment 3. This experiment was carried out with data points in R^3. Here data sets were generated from two classes A_6 and A_7 using triangular distribution in R^3, where

$$A_6 = [0,1] \times [0,1] \times [0,1]$$

and

$$A_7 = [1,2] \times [0,1] \times [0,1].$$

Table 3 presents the results for this experiment. It can be seen from Table 3 that the average error $e_n(\omega)$ is going towards Bayes error probability $e(\delta)$ as the number of data points in the given sample increases.

Experiment 4. The theoretical results presented in this paper are based on pattern classes which are bounded. On the other hand, pattern classes that are generated using normal distribution are not bounded. We have applied the proposed method on unbounded classes also to check whether it provides meaningful results experimentally. Table 4 shows the experimental results with data sets that are generated from two classes A_8 and A_9 in R^2 using normal distribution. The mean values for the two classes

Table 3. Results with 2 class problem in 3-D using triangular distribution

δ Value	No. of points	Average Error $e_n(\omega)$	Bayes Error Probability $e(\delta)$
	500	0.30666	
0	600	0.20800	0
	700	0.10680	

Table 4. Results with 2 class problem in 2-D using normal distribution

δ Value	No. of points	Average Error $e_n(\omega)$	Bayes Error Probability $e(\delta)$
	1000	0	
6.5	1100	0	0.00576
	1200	0	
	1000	0.10443	
6	1100	0.00640	0.00134
	1200	0.00708	

are taken to be μ_1 and μ_2, where

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \mu_{2\delta} = \begin{pmatrix} \delta \\ 0 \end{pmatrix}; \delta = 6, 6.5.$$

The variance for both the classes are taken to be unity. The covariance between the variables is taken to be zero. The *a priori* probability for the two classes are taken to be equal to $\frac{1}{2}$. The Bayes error probability $e(\delta)$ is given by

$$e(\delta) = \int_{\delta/2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-1/2)\, dx.$$

It can be seen from Table 4 that the average error $e_n(\omega)$ is going towards Bayes error probability $e(\delta)$ for $\delta = 6$ as the number of data points in the given sample increases. Note that the e_n values are all zero for $\delta = 6.5$ and higher, although we have a non-zero Bayes error probability for those δ values. But note that the regions corresponding to the clusters generated by the proposed method is bounded whereas the classes provided by the Bayes are unbounded

7. CONCLUSION AND DISCUSSION

In this paper, a clustering technique is presented which extracts clusters by finding the valley regions in the feature space. We have also provided the theoretical formulation of the technique. It is shown theoretically that the proposed clustering technique provides the same result as that of a Bayes classifier as the number of data points goes to infinity under a smooth assumption.

Different clustering techniques use different clustering criterions to obtain different partitionings of a given data set. Thus the validity of the obtained clusters is to be judged on some "suitable" basis to ensure the effectiveness of the clustering technique. Note that Bayes classifier provides the regions corresponding to different classes with minimum error probability. In this work, a clustering technique is proposed and its performance is judged both theoretically and experimentally with respect to that of Bayes classifier. Note that Bayes classifier handles the problem of supervised pattern recognition whereas conventional clustering is viewed as a problem of unsupervised pattern recognition. It has been shown, both theoretically and experimentally, that under a "smooth" condition the performance of the proposed clustering technique tends to that of Bayes classifier as the number of data points in a given data set increases. Note that the proposed clustering technique, unlike Bayes method, starts with no knowledge about the pattern classes or clusters present in a given data set. The proof given for convergence is based on certain mild assumptions[18] on the density function.

The proposed technique basically finds the valley regions in the multidimensional feature space. A similar approach is adopted in image processing for segmenting the image using gray level intensities with the help of histogram thresholding.[20–23] But the aspect of judgment of the "quality" of the classes obtained in such a case with respect to that of Bayes classifier is not usually attempted. In this work, the problem is treated analytically and it is shown, both theoretically and experimentally, that the proposed clustering technique, which uses the valley regions of the data to generate clusters, would provide the same number of clusters as the number of classes obtained using Bayes classifier and each such cluster would correspond to exactly one unique Bayes class, as the number of data points goes to infinity.

This article also presents a method of finding "valley regions" in multivariate histogram, the literature on which is inadequate till date to the best of knowledge of the authors. Though the computational complexity for finding such regions is exponential with respect to the number of dimension, the procedure has been found to be theoretically good in providing the Bayes classes for sufficiently large number of observations (n). The problem of reducing the computational complexity for finding the clusters in higher dimension has not been attempted in this work.

The article also emphasizes the need for validating clustering techniques with the help of Bayes classifier. Artificial data set following a known mixture density function may be generated and the performance of a clustering technique may be judged by comparing the number of misclassified points of the clustering technique with that of the Bayes classifier. For a proper judgment of the applicability of any clustering technique, such a comparison should be attempted on a large number of data sets with varying sizes, varying number of classes and varying mixture density function for obtaining a meaningful conclusion.

The proposed technique assumes that the boundary between any two classes must belong to the valley region. But note that we may have data distribution where the assumption is not valid. For example in Experiment 1, if δ value is taken to be less than 3, then the Bayes decision boundary may correspond to a local maxima in the mixture density function. Similarly one can cite several other examples where the assumption is not valid. But note that in many real life data sets, the density of the data points in the vicinity of the boundary between in any two classes is less compared to that in the core regions of the classes. Thus one can use the proposed technique in finding the valley regions of the feature space and treat them as the boundary between classes. The clusters thus obtained are expected to correspond to the original classes present in the feature space.

The proposed clustering technique has been defined to be applicable for bounded pattern classes only. But it is found experimentally (Table 4) that the technique provides good results in the case of unbounded pattern classes also.

REFERENCES

1. M. R. Anderberg, *Cluster Analysis for Application*. Academic Press, New York (1973).
2. T. Tou Julius and C. Gonzalez Rafael, *Pattern Recognition Principles*. Addison-Wesley, Reading, Massachusetts (1974).
3. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. Wiley, New York (1973).
4. K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, New York (1972).
5. A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, New Jersey (1988).
6. P. A. Devijver and J. Kittler, *Pattern Recognition: A statistical Approach*. Prentice-Hall International, Hemel Hemstead, Hertfordshire, England (1982).
7. R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. Prentice-Hall, New Jersey (1982).
8. H. Spath, *Cluster Analysis Algorithms*. Ellis Horwood, Chichester, U.K. (1980).
9. C. A. Ankerbrandt, B. P. Buckles and F. E. Petry, Scene recognition using genetic algorithms with semantic nets, *Pattern Recognition Lett.* **11**, 285–293 (1990).
10. C. A. Murthy and N. Chowdhury, In search of optimal clusters using Genetic Algorithms, *Pattern Recognition Lett.* **17**, 825–832 (1996).
11. S. Z. Selim and M. A. Ismail, K-MEANS type algorithms: A generalized convergence theorem and characterization of local optimality, *IEEE Trans.* **PAMI-6(1)**, 81–87 (1984).
12. C. A. Murthy, On consistent estimation of class in R^2 in the context of cluster analysis, Ph.D. Thesis. ISI, Calcutta (1989).
13. D. P. Mandal and C. A. Murthy, Selection of alpha for alpha-hull and formation of fuzzy alpha-hull in R^2, *Int. J. Uncertainty Fuzziness Knowledge based Systems* **4**, 401–417 (1995).
14. D. P. Mandal, C. A. Murthy and S. K. Pal, Determining the shape of a pattern class for sampled points: Extension to R^n, *Int. J. General Systems* (to appear) (1995).
15. D. P. Mandal and C. A. Murthy, Selection of alpha for alpha-hull in R^2, *Pattern Recognition* (to appear) (1995).
16. D. Chaudhuri, C. A. Murthy and B. B. Chaudhuri, Finding a subset of representative points in a data set, *IEEE SMC* **9**, 1416–1424 (1994).
17. N. Chowdhury and C. A. Murthy, *Finding the natural grouping in a data set* (accepted) (1994).
18. D. Chaudhuri, B. B. Chowdhury and C. A. Murthy, A data driven procedure for density estimation with some applications, *Pattern Recognition* (to appear).
19. T. M. Cover and P. E. Hart, Nearest neighbor pattern classification, *IEEE Trans. Inform. Theory* **13**, 21–67 (1967).
20. V. Ryzin, On a histogram method of density estimation, *Comm. Statist.* **2**, 493–506 (1973).
21. R. D. Reiss, Approximate distribution of the maximum deviation of histograms, *Metrika* **25**, 9–26 (1978).
22. J. A. Richards, *Remote Sensing Digital Image Analysis: An Introduction*. Springer, Berlin (1986).
23. W. L. G. Koontz and K. Fukunaga, A nonparametric valley seeking technique for cluster analysis, *IEEE Trans. Comput.* **C-21**, 171–178 (1972).

About the Author — NIRMALYA CHOWDHURY was born in 1963 in Calcutta, India. He received the B.Sc. (honours) degree in Physics in 1983, the B.Tech (Radiophysics and Electronics) degree in 1987 and the M.Tech (Computer Science and Engineering) degree in 1990 from Calcutta University, Calcutta, India. He has submitted his thesis for the Ph.D. degree in August 1996 at Jadavpur University, Calcutta, India. His research interests include pattern recognition, image processing, hard and fuzzy clustering, neural networks and genetic algorithms. He is currently working as a Lecturer in the Department of Electrical Engineering at Ramakrishna Mission Shilpapitha, Belgharia, Calcutta, India. He is a member of Indian Society for Technical Education and an associate member of The Institution of Engineers (India).

About the Author — C. A. MURTHY was born in 1958 in Ongole, India. He received the B.Stat. (Hons.) degree in 1979, the M.Stat. degree in 1980 and the Ph.D. degree in 1989 from Indian Statistical Institute, Calcutta, India. His research interests include pattern recognition, image analysis, fuzzy sets, cluster analysis, fractals, neural networks and genetic algorithms. He is currently working as an Associate Professor in the Machine Intelligence Unit at the Indian Statistical Institute, Calcutta, India. He is a member of Indian Society for Fuzzy Mathematics and Information Processing, and Indian Unit for Pattern Recognition and Artificial Intelligence.