A NONPARAMETRIC VALLEY-SEEKING TECHNIQUE FOR CLUSTER ANALYSIS

Warren Lewis Gordon Koontz
and
Keinosuke Fukunaga

School of Electrical Engineering
Purdue University
West Lafayette, Indiana U.S.A.

Abstract

The problem of clustering multivariate observations is viewed as the replacement of a set of vectors with a set of labels and representative vectors. A general criterion for clustering is derived as a measure of representation error. Some special cases are derived by simplifying the general criterion. A general algorithm for finding the optimum classification with respect to a given criterion is derived. For a particular case, the algorithm reduces to a repeated application of a straightforward decision rule which behaves as a valley-seeking technique. Asymptotic properties of the procedure are developed. Numerical examples are presented for the finite sample case.

I. Introduction

It is not difficult to imagine a collection of objects whose members can be classified into two or more categories simply on the basis of their observable characteristics. It is not always necessary to rely on a similar collection of labeled objects as a basis for classification. For example, biological taxonomists have classified living things into a large number of meaningful categories. Yet at no time in history did any plant or animal bear a label. Rather, categories have been established without supervision.

Recently, methods for automatic unsupervised classification, or clustering, have been proposed. A machine algorithm for clustering can be a valuable tool in

i) pattern recognition - Often, a training set of labeled objects is difficult or impossible to obtain. Further, a known class of objects may contain unknown subclasses,

and

ii) statistical analysis - Cluster analysis may be used to expose the detailed structure of a large volume of data.

We will present and discuss a family of clustering algorithms.

Our approach involves the use of a clustering criterion. This criterion assigns a numerical value to every possible classification of the objects. Meaningful classifications are assumed to correspond to extreme values of the criterion. The optimum classification in the sense of a given criterion is determined by means of a clustering algorithm. An efficient clustering algorithm is necessary because an exhaustive check of all possible classifications is usually impractical. Thus, for our purposes, the clustering problem consists of two basic elements:

1) definition of a clustering criterion

and

11) construction of a clustering algorithm.

The idea of using a clustering criterion is not new. Many procedures reported in Ball's survey (1) are based on criteria. Friedman and Rubin (2) present a class of criteria and discuss the property of transformation invarlance. Fukunaga and Koontz (3) show conditions where the criteria of (2) become equivalent to a simpler criterion. Watanabe (4) proposes a criterion, which he calls cohesion, which can detect more subtle relationships among objects than palrwlse similarities.

Presently, no universal clustering criterion has been defined. This is simply a consequence of the lack of a precise mathematical definition of a cluster. That is, the clustering problem is one whose solution cannot be characterized in a definite way. Thus, in order to derive a mathematical criterion, we must postulate a rigorous definition or clustering. This postulate can then be tested by experiments with objects whose class structure is known and well defined.

The remainder of this paper consists of four sections and a summary. In the next section, (section II) we present our characterization of the clustering problem and compare it with other notions. We will then use this characterization to derive a criterion. In the following section we will state and discuss a general algorithm for finding the classification which extremizes our criterion. Section IV concerns the asymptotic behavior of the procedure, i.e., what happens when the number of objects is very large. Results of computer experiments are given in section V.

II. A Clustering Criterion

The criterion derived in this section is based on the notion that information is lost when objects are represented only by class labels. Suppose that each member of a collection of N objects is represented by an L-dlmenslonal vector. Then the set of N vectors, $[X_1 \ldots X_N\}$, contains all of the available Information concerning the objects. The clustering operation replaces this set of vectors with a set of labels, $\{w_1, \ldots, w_N\}$. The i-th label, w_1, is an integer between 1 and M (M < N), and denotes the class to which X_1 is assigned. The label set contains less information about the objects than does the vector set. Therefore, clustering is viewed here as a data reduction algorithm which destroys information.

This loss of information can be viewed as the representation error committed by replacing $\{\underline{X}_1, \ldots, \underline{X}_N\}$ with $\{\omega_1, \ldots, \omega_N\}$. A quantitative measure of this error can serve as a clustering criterion to be minimized. There are at least two ways to derive a numerical measure of representation error. The method which has been used most often in the past is to measure the error committed by using a representative vector, C(w1), *aa* an estimate of X_i. An error vector can be defined as

$$\underline{e}_i = \underline{X}_i - \underline{C}(\omega_i), \qquad [1]$$

and a cumulative error matrix is given by

$$\underline{W} = \sum_{i=1}^{N} \underline{e}_i \underline{e}_i^T \qquad [2]$$

If $\underline{C}(j)$ is the mean of class j, i.e.,

$$\underline{C}(j) = \frac{1}{N_j} \sum_{\omega_i = j} \underline{X}_i, \qquad [3]$$

where N_i. is the population of class j, then W is the total intragroup scatter matrix. Several criteria which are functions of W are discussed in (2) and (3).

An alternate definition of representation error is used in the present development. We will concern ourselves with the error committed in estimating distances between pairs of vectors. Let $d_X(\underline{X}_i, \underline{X}_j)$ be the euclidean distance between $\underline{X}_i$ and $\underline{X}_j$,

$$d_X(\underline{X}_i, \underline{X}_j) = \left[\sum_{\ell=1}^{L} (X_{i\ell} - X_{j\ell})^2 \right]^{1/2} \qquad [4]$$

Further, let $d_\omega(\omega_i, \omega_j)$ be a suitably defined metric of interclass distances. For example

$$d_\omega(\omega_i, \omega_j) = d_X[\underline{C}(\omega_i), \underline{C}(\omega_j)], \qquad [5]$$

or

$$d_\omega(\omega_i, \omega_j) = \begin{cases} D > 0 & \omega_i \neq \omega_j, \\ 0 & \omega_i = \omega_j \end{cases} \qquad [6]$$

Then a measure of distance representation error, which will be used as a clustering criterion, is

$$J = \sum_{i=1}^{N} \sum_{i=1}^{N} f(\underline{X}_i, \underline{X}_j) \left[d_\omega(\omega_i, \omega_j) - d_X(\underline{X}_i, \underline{X}_j) \right]^2 \qquad [7]$$

where f(Xi, Xi) is a set of weighting coefficients. This kind of criterion is often used to measure mapping error and clustering is a kind of mapping. However, some special considerations are important in its use as a clustering criterion. First of all, not all of the distances are euclidean. A more important point, however, is the fact that the w_i's are variables and the X_i's are fixed. Due to the discrete and unordered nature of the w_i's, ordinary gradient methods cannot be used to minimize J.

Criteria of the same form as J have been used in hierarchical clustering. In hierarchical clustering, objects are classified according to a diverging tree structure. A tree metric is defined which numerically defines the distance between two objects according to their position on the tree. The degree of fit between the a measure of validity of the classification tree (5,6).

The general criterion, J is too cumbersome to use in practice. The summation contains N^2 terms in general. Therefore, we would like to assign zero weight to most of the terms. Suppose f satisfies

$$f(\underline{X}_i, \underline{X}_j) = 0, \quad d_X(\underline{X}_i, \underline{X}_j) > R > 0 \qquad [8]$$

If R is sufficiently small, then J can be approximated as

$$J \doteq \sum_{i=1}^{N} \sum_{j=1}^{N} f(\underline{X}_i, \underline{X}_j)\, d_\omega^2(\omega_i, \omega_j) \triangleq J_1 \qquad [9]$$

A specific example of J_1 follows when f is defined by

$$f(\underline{X}_i, \underline{X}_j) = \begin{cases} 1, & d_X(\underline{X}_i, \underline{X}_j) \leq R, \\ 0, & d_X(\underline{X}_i, \underline{X}_j) > R, \end{cases} \qquad [10]$$

$$\triangleq f_R[d_X(\underline{X}_i, \underline{X}_j)] .$$

Since f_R is symmetric with respect to $\underline{X}_i$ and $\underline{X}_j$ and since $d_\omega(\omega_i, \omega_i) = 0$, i=1,...,N (a property of any metric), we can write J_1 as

$$J_1 = 2 \sum_{j<i} f_R[d_X(\underline{X}_i, \underline{X}_j)] d_\omega^2(\omega_i, \omega_j) \triangleq 2 J_{1R} \qquad [11]$$

where the notational equivalence

$$\sum_{j<i} = \sum_{i=2}^{N} \sum_{j=1}^{i-1} \qquad [12]$$

is implied.

JIP assigns a nonzero penalty for each pair of vectors closer together than R and classified into different classes.

If [6] is taken as the definition of d , the following *special* cases of Ji and J_{IR} result:

$$J_1 = \sum_{i=1}^{N} \sum_{j=1}^{N} f(\underline{X}_i, \underline{X}_j)\, D^2[1 - \delta(\omega_i, \omega_j)], \qquad [13]$$

$$\triangleq D^2 J_2 ,$$

$$J_{1R} = \sum_{j<i} f_R[d_X(\underline{X}_i, \underline{X}_j)]\, D^2[1 - \delta(\omega_i, \omega_j)] , \qquad [14]$$

$$\triangleq D^2 J_{2R} ,$$

where

$$\delta(\omega_i, \omega_j) = \begin{cases} 1 & \omega_i = \omega_j, \\ 0 & \omega_i \neq \omega_j. \end{cases} \qquad [15]$$

J2R is the simplest criterion we will derive. It is equal to the total number of distinct pairs of vectors separated by a distance less than R and assigned to different classes. We will sometimes refer to J2R as the fixed neighborhood penalty rule. The remainder of this paper mainly concerns J_{2R}.

The following properties of J_{2R} support its use as a clustering criterion.

1) Comnuttion For *ifiction*

J_{2R} is evaluated by a counting process rather than by complex calculations. Since the vectors are fixed, the neighboring pairs need be determined only once.

ii) <u>Storage</u> When R is sufficiently small, the number of neighboring pairs of vectors is moderate. The storage requirement is governed primarily by this number,

iii) <u>Classification</u> Contributions to JOD come from pairs of vectors near the boundaries separating classes. Thus, It is preferable for the boundary to pass through a region of low vector concentration. This kind of classifica tion is quite reasonable when there is no supervision available.

At this point, the reader may wonder if a K nearest neighbor penalty rule can be defined. The answer is yes, since we can write

$$J_{?K} = \sum_{i=1}^{N} \sum_{j=1}^{N} f_K(\underline{X}_i,\underline{X}_j)[1-\delta(\omega_i,\omega_j)] \qquad [16]$$

where

$$f_K(\underline{X}_i,\underline{X}_j) = \begin{cases} 1, \text{ if } \underline{X}_j \text{ is one of the K nearest neighbors of } \underline{X}_i, \\ 0, \text{ otherwise.} \end{cases} \qquad [17]$$

Notice that fk is not symmetric. Although the K nearest neighbor penalty rule has a valuable counterpart in supervised pattern recognition, it is unsuitable in the present case. The K nearest neighbor rule does not favor one region over another because of density. Therefore, it may well prescribe a boundary through a mode in the vector distribution.

At this point, we have a family of nonparametric criteria with three levels of complexity. The parent criterion, J, is the most complex and is in the closest accord with our original concept of clustering (distance preservation). Its descendants are J1 followed by J2, with special cases J_{IR} and J_{2R}. Criteria at the J1 level are more general in that they allow the penalty to be class dependent, but J_{2R} is easier to implement and admits an interpretation which seems very suitable. Unfortunately, criteria of the J1 level and below have an absolute minimum of zero when all vectors are assigned to the same class. This is not a serious problem in practice because there will be local minima corresponding to more interesting classifications. The degenerate case is easily detected.

We have not specified how to choose either the number of classes, M, or the region size, R. We have no rigorous theory to rely on here, and we can only offer suggestions based on experimental results. Therefore, we postpone discussion of these points until section V.

III. The Clustering Algorithm

The algorithm for finding the optimum assignment with respect to a given criterion is the second essential ingredient of clustering. Although clustering need not take place in real time, there are still practical constraints which rule out inefficient procedures, such as exhaustive searching. We have made use of a general type of algorithm. This algorithm can be applied to a wide variety of criteria, but in the special cases of J_{2R} it becomes particularly easy to implement.

Consider a criterion of the form

$$J = J(\omega_1,\ldots,\omega_N;\ \underline{X}_1,\ldots,\underline{X}_N), \qquad [18]$$
$$= J(\underline{\Omega};\ \underline{X}),$$

where

$$\underline{\Omega} = [\omega_1,\ldots,\omega_N]^T, \qquad [19]$$

and

$$\underline{X} = [\underline{X}_1^T \ldots \underline{X}_N^T]^T. \qquad [20]$$

The assignment, $\underline{\Omega}$, is variable and the configuration, $\underline{X}$, is fixed. We are seeking an assignment, $\underline{\Omega}^*$, such that

$$J(\underline{\Omega}^*,\underline{X}) = \min_{\underline{\Omega}} J(\underline{\Omega};\underline{X}). \qquad [21]$$

Because of the discrete and unordered nature of $\underline{\Omega}$, ordinary gradient methods cannot be used. Still, it is possible to specify an iterative search based on first order variations in J with respect to $\underline{\Omega}$. Let $\underline{\Omega}(k)$ the assignment at the kth step. If the rth vector is reclassified from its present class, $\omega_r(k)$, to class s, then the change in J, $\Delta J(r,s,k)$, is given by

$$\Delta J(r,s,k) = J[\omega_1(k),\ldots,\omega_{r-1}(k),s,\omega_{r+1}(k),\ldots,\omega_N(k);\underline{X}] - J[\underline{\Omega}(k);\underline{X}]. \qquad [22]$$

The succeeding classification of $\underline{X}_r$, $\omega_r(k+1)$ should be the one which yields the most negative change in J. Therefore, the following clustering algorithm is proposed.

Step 1: Choose an initial classification, $\underline{\Omega}(0)$.

Step 2: Having determined the kth classification, calculate $\Delta J(r,s,k)$ for r=1,...,N and s=1,...,M.

Step 3: The k+1st classification is determined by

$$\Delta J(r,\omega_r(k+1),k) = \min_{s} \Delta J(r,s,k), \quad r=1,\ldots,N \qquad [23]$$

Ties involving $w_r(k)$ are resolved in its favor. Other ties are resolved arbitrarily.

Step 4: If any vector is placed in a new class, return to Step 2 and repeat. Otherwise, stop.

Note that all computation occurs in step 2 and the vectors are reclassified simultaneously in step 3.

There is no guarantee that this algorithm will converge. Even if it does, there is little we can say about the strength of the minimum obtained. Fortunately, empirical evidence seems to favor this procedure. Fukunaga and Koontz

This procedure, which follows from application of the general algorithm to a specific criterion, J2R, is a valley seeking technique. To see this, consider vectors along the boundary separating class S1 from class S2 at the kth iteration. Suppose there is a heavier concentration of vectors on the $s_?$ side of the boundary. Then vectors near the boundary are reclassified into class S?. Hence, the boundary moves into the region previously assigned to class sj. Therefore, the boundary moves away from the higher concentrations and toward valleys in the distribution.

Two kinds of difficulty may arise when the fixed neighborhood decision rule is used. First of all, the algorithm may get stuck with the boundary passing through a region of relatively sparse population when better boundaries exist. Secondly, the boundary may diverge, leaving all of the vectors in a single class. Both of these difficulties are combatted by altering the initial assignment, (0), and adjusting the control parameter, R.

The clustering algorithm and the clustering criterion together make up a clustering procedure. The clustering procedure has become the iterative application of the fixed neighborhood decision rule. We can easily apply it to numerical examples, and we do this in section V. First, however, let us see how the procedure behaves when N is very large.

IV. Asymptotic Behavior

The performance of the clustering algorithm developed in sections II and III can be studied analytically when N is very large. In this section, we will derive the asymptotic version of D_{2R} and discuss its properties. The asymptotic properties provide some insight into the general behavior of our procedure. They also suggest how the procedure can be expected to perform with finite data sets.

Let us first rewirte the expression for D_2R, normalizing bv a factor of 1/N.

$$D_{2R}(r,s) = \frac{1}{N}\sum_{i=1}^{N} f_R[d_X(\underline{X}_r,\underline{X}_i)] - \frac{1}{N}\sum_{i=1}^{N} f_R[d_X(\underline{X}_r,\underline{X}_i)]\delta(s,\omega_i) \qquad [26]$$

The first term of [26] is independent of s. Therefore, choosing the minimum of D_{2R} is equivalent to choosing the maximum of

$$D^*_{2R}(r,s) = \frac{1}{N}\sum_{i=1}^{N} f_R[d_X(\underline{X}_r,\underline{X}_i)]\delta(s,\omega_i) \qquad [27]$$

Let Q_s be the set of all vectors assigned to class s. Then, as N becomes large, D^*_{2R} approaches an integral.

$$D^*_{2R}(r,s) \rightarrow \int_{Q_s} f_R[d_X(\underline{X}_r,\underline{X})]p(\underline{X})\,d\underline{X} \triangleq D^*_{2R}(\underline{X}_r,s) \qquad [28]$$

where $p(\underline{X})$ is the mixture probability density function of the $\underline{X}_i$'s. Let $\underline{Y}$ be an arbitrary point to be reclassified and let $S_R(\underline{Y})$ be the set of all vectors separated from $\underline{Y}$ by a distance less than R. Then

$$D^*_{2R}(\underline{Y},s) = \int_{Q_s \cap S_R(\underline{Y})} p(\underline{X})\,d\underline{X} \qquad [29]$$

The behavior of the decision rule corresponding to [29] is easily illustrated when the dimension, L, is two. Figure 1 shows a region around the boundary separating classes S_1 and S2. For the value of R shown, $\underline{Y}_1$ clearly remains in class S_1. However, if the probability mass within R of Y_2 and to the right of the boundary is larger than that to the left of the boundary, then Y2 is reassigned to class S_2. If [29] reassigns no vectors, then the boundary is said to be stationary.

If R is sufficiently small, we can characterize a stationary boundary rather nicely. Figure 2 shows a small region about a point on the boundary between classes s_1 and s_2. The boundary has unit normal vector w and $\nabla p(\underline{Y})$ is the gradient of the mixture density evaluated at Y, i.e.,

$$\nabla p(\underline{Y}) = \left[\frac{\partial p}{\partial X_1}\bigg|_{\underline{X}=\underline{Y}} \cdots \frac{\partial p}{\partial X_L}\bigg|_{\underline{X}=\underline{Y}} \right]^T \qquad [31]$$

Since R is small, p($\underline{X}$) can be approximated closely with a truncated Taylor expansion about $\underline{Y}$. The decision rule then becomes

$$V_1\, p(\underline{Y}) + [\nabla p(\underline{Y})]^T \int_{Q_{s_1} \cap S_R(\underline{Y})} (\underline{X}-\underline{Y})\, d\underline{x} \underset{S_2}{\overset{S_1}{\gtrless}} V_2 p(\underline{Y}) + [\nabla p(\underline{Y})]^T \int_{Q_{s_2} \cap S_R(\underline{Y})} (\underline{X}-\underline{Y})\, d\underline{X} \qquad [32]$$

where V. is the volume of Q. S_R (Y) and the superscript T denotes transposition. Agein noting that R is small, we assume that the boundary splits S_R(Y) into two L dimensional hemispheres so that $V_1 = V_2$. The integrals in [32] are given by

$$\int_{Q_{s_1} \cap S_R(\underline{Y})} (\underline{X}-\underline{Y})\, d\underline{X} = -\alpha \underline{w} \qquad [33]$$

and

$$\int_{Q_{s_2} \cap S_R(\underline{Y})} (\underline{X}-\underline{Y})\, d\underline{X} = \alpha \underline{w} \qquad [33]$$

where

$$\alpha = \frac{R^{L+1}}{L+1}\, \frac{\pi^2}{\Gamma(\frac{L}{2}+1)}\, . \qquad [34]$$

Thus, the final form of the decision rule is

$$-\alpha[\nabla p(\underline{Y})]^T \underline{w} \underset{s_2}{\overset{s_1}{\gtrless}} 0. \qquad [35]$$

Suppose the left hand side of [35] is positive. Then Y will be assigned to s_1. Further, all vectors within a small neighborhood of the boundary will also go to s_1. Thus the boundary shifts to the right (see Fig. 2). Similarly, if the right hand side of [35] is negative, the boundary moves to the left. The condition for stationarity of the boundary is

$$\nabla^T p(\underline{Y})\, \underline{w} = 0. \qquad [36 1$$

A final boundary between two classes must be stable as well as stationary. This means that if the boundary is perturbed it must not tend to move farther away from the stationary point. We can establish a condition for unstability as follows. Figure 3 is an exagerated illustration of a small perturbation of a stationary boundary. The vector Y' is a point on the new boundary such that

$$\underline{Y}' - \underline{Y} = \beta\, \underline{w}', \quad \beta > 0. \qquad [37]$$

where w is the new unit normal vector. If the component of ∇p(Y) along w is negative, then the boundary will tend to move farther away from the stationary position. Hence the boundary is unstable if

$$[\nabla p(\underline{Y}')]^T w' < 0 \qquad [38]$$

for any small perturbation.

We can express $\nabla p(\underline{Y}')$ using a Taylor series about $\underline{Y}$ as

$$\nabla p(\underline{Y}') \doteq \nabla p(\underline{Y}) + [\nabla^2 p(\underline{Y})]^T (\underline{Y}'-\underline{Y}), \qquad [39]$$

where $\nabla^2 p(\underline{Y})$ is a matrix

$$[\nabla^2 p(\underline{Y})]_{ij} = \frac{\partial^2 p(X)}{\partial X_i\, \partial X_j}\bigg|_{\underline{X}=\underline{Y}} \qquad [40]$$

Using [37] and [39] in [38] we can write

$$[\nabla p(\underline{Y}')]^T \underline{w}' \doteq [\nabla p(\underline{Y})]^T \underline{w}' + \underline{w}'^T [\nabla^2 p(\underline{Y})]\, \underline{w}'\, . \qquad [41]$$

Suppose $\nabla^2 p(\underline{Y})$ is negative semidefinite. Then the second term of [41] is nonpositive for all $\underline{w}'$. The direction of $\underline{w}'$ is arbitrary so that we can choose $\underline{w}'$ such that the first term of [41] is negative. Hence, if $\nabla^2 p(\underline{Y})$ is negative semidefinite, then for some $\underline{w}'$ [38] holds and the boundary is therefore unstable.

Tn conclusion, the final boundary must satisfy two conditions.

i) The component of the gradient of the density normal to the boundary must be zero.

ii) The boundary may not pass through regions where $\nabla^2 p$ is negative seraidefinite.

This development shows that our algorithm leads to reasonable classifications in the asymptotic case. Hopefully, it also provides insight into the behavior of the algorithm in the finite sample case as well.

V. Examples

The algorithm has been tested on artifically generated bivariate data. There is no additional difficulty in the multivariate case.

The value of R has considerable effect on the performance of the algorithm. We found that the procedure works best when R is such that the number of distances less than R is 10 to ?0 times the sample size.

The choice of M is more difficult. In one case, a large value of M resulted in most of the vectors being placed in one of two classes, but we cannot guarantee that this would always be the result.

Figure 4 show the results of one example with M=2. The initial boundary is random. Note that the data are not linearly separable.

The number of iterations required in the experiments ranged from 4 to 10. Total computation time was-under 10 seconds on a CDC 6500.

VI. Summary

A clustering procedure consists of a criterion and an algorithm. We have developed a general clustering procedure of which the fixed neighborhood decision rule is a special case. The asymptotic behavior of the procedure was studied and computer experiments testified to its practical value.

The procedure has been shown to be suitable even for non ellipsoidal clusters. It has modest storage requirements and the computational loop, which involves only counting, is very rapid.

References

(1) G.H. Ball, "Data analysis in the social sciences: what about the details?", 1965 Fall Joint Computer Conf.. AFIPS Proc, vol. 27, pt. 1, Washington, D.C.: Spartan, 1965, pp. 533-559.

(2) H.P. Friedman and J. Rubin, "On some invariant criteria for grouping data", Amer. Stat. Assoc. J., vol. 62, pp. 1159-1178, December, 1967.

(3) K. Fukunaga and W.L.G. Koontz, "A criterion and an algorithm for grouping data", IEEE Trans, on Computers, vol. C-19, No. 10, pp. 917-923, October, 1970.

(4) M.S. Watanabe, Knowing and Guessing, New York: Wiley, 1969, Ch. 8.

(5) S.C. Johnson, "Hierarchical clustering schemes", Psychometrika, vol. 32, pp. 241-254, 1967.

(6) J.A. Hartigan, "Representation of similarity matrices by trees", Amer. Stat. Assoc. J., vol. 62, pp. 1140-1158, December, 1967.

(7) K. Fukunaga and W.L.G. Koontz, "Application of the Karlunen-Loeve expansions to feature selection and ordering", vol. C-19, no. 4, pp. 311-318, April, 1970.

(8) E.H. Ruspini, "A new approach to clustering," Information and Control, vol. 15, pp. 22-32, 1969.

(9) F.J. Rohlf, "Adaptive hierarchical clustering schemes", Systematic Zoology, vol. 9, no. 1, pp. 58-82, March 1970.

(10) R. Gnanadesikan and M.B. Wilk, "Data analytic methods in multivariate statistical analysis", Multivariate Analysis, Vol. II, New York: Academic Press, 1969.

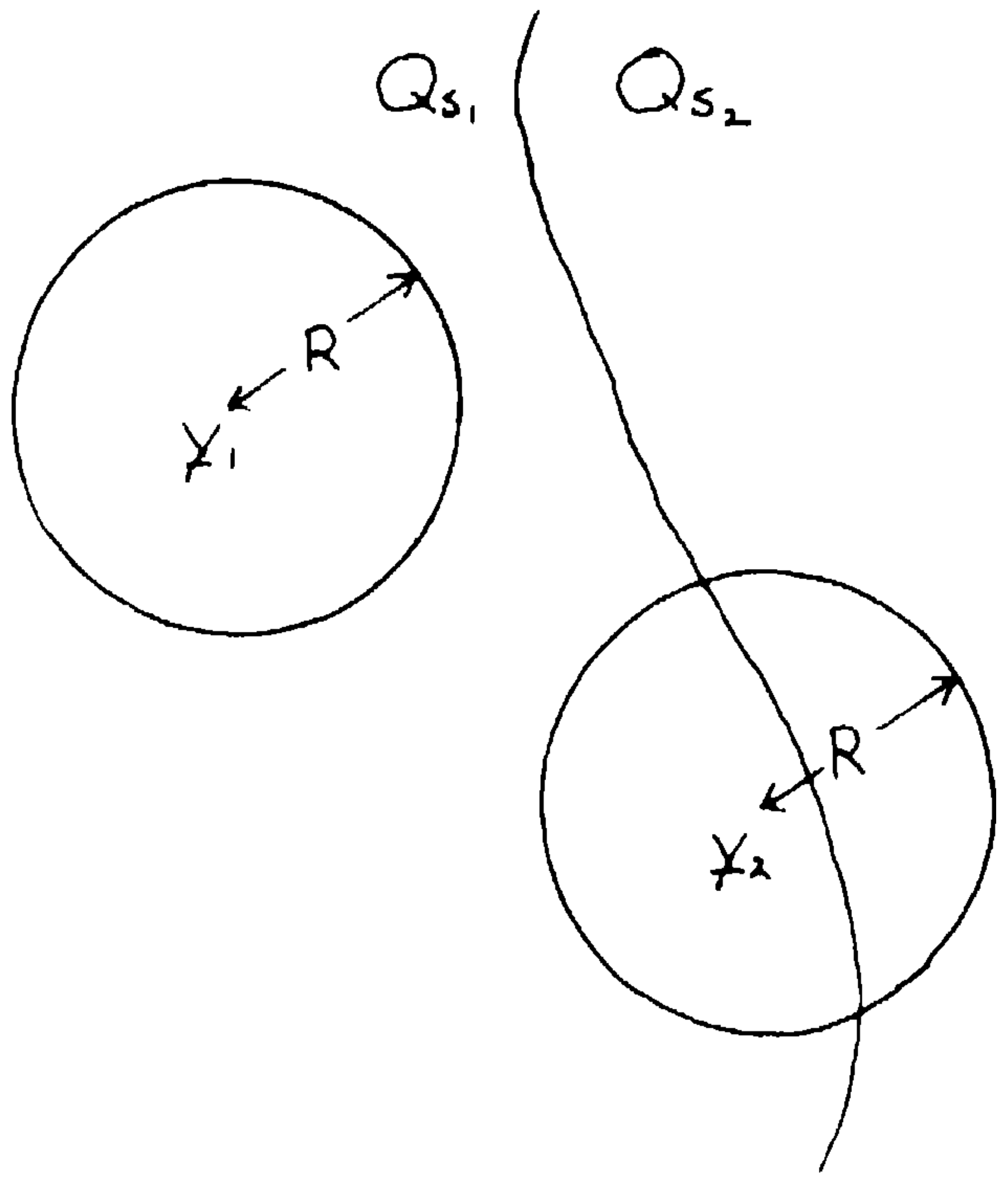

Figure 1.
Asymptotic fixed neighborhood decision rule

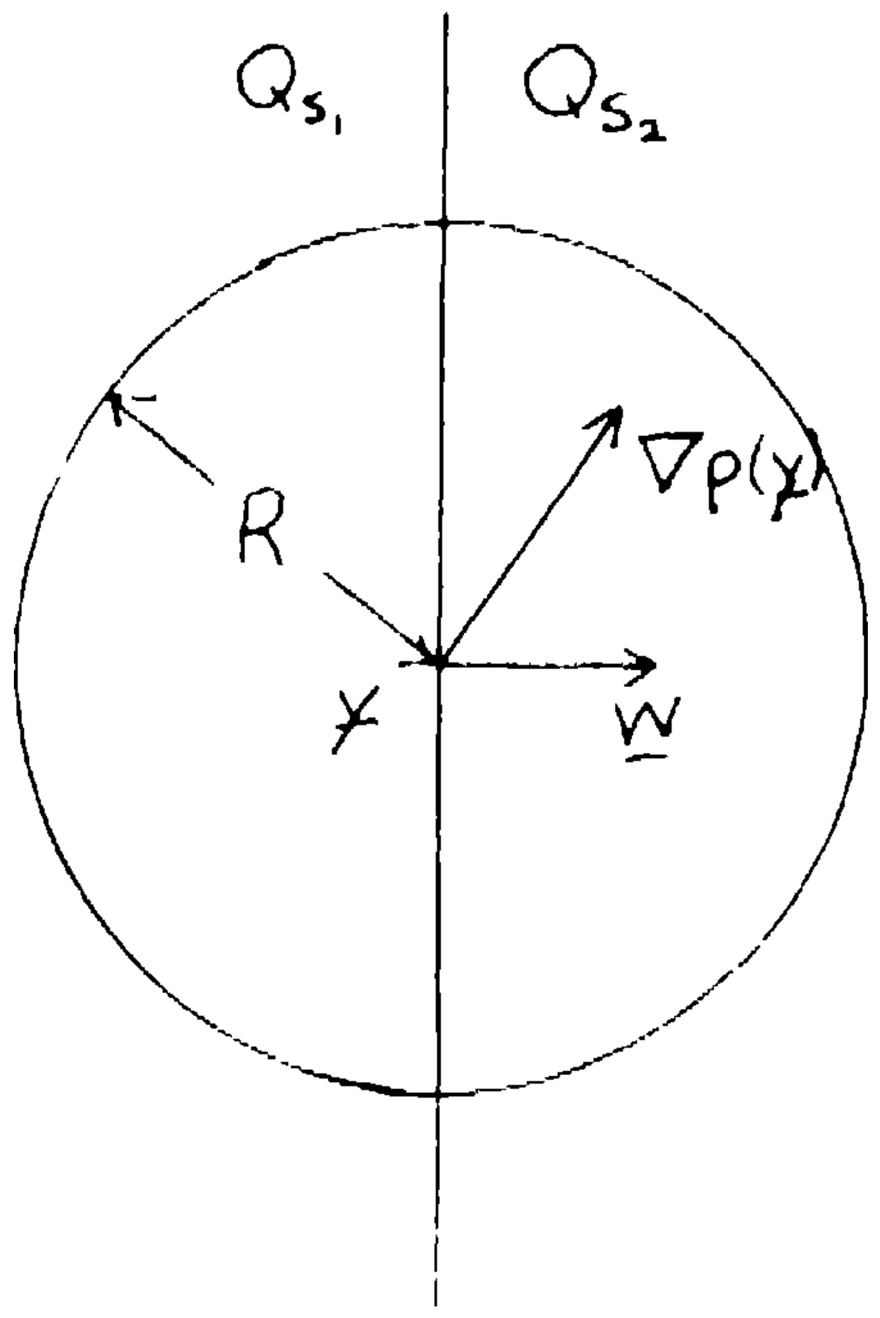

Figure 2.
Classification of a boundary point.

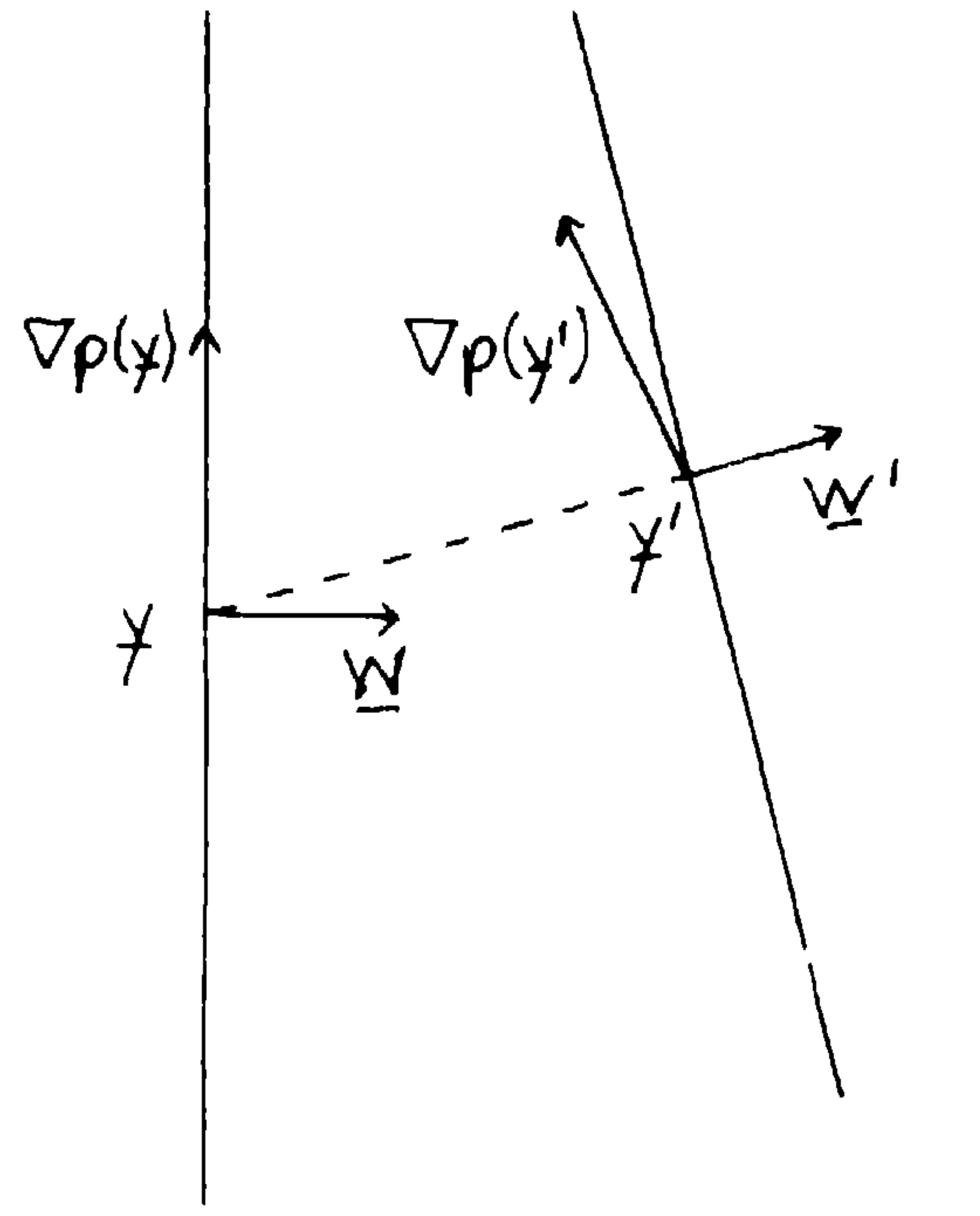

Figure 3.
Perturbation of a stationary boundary

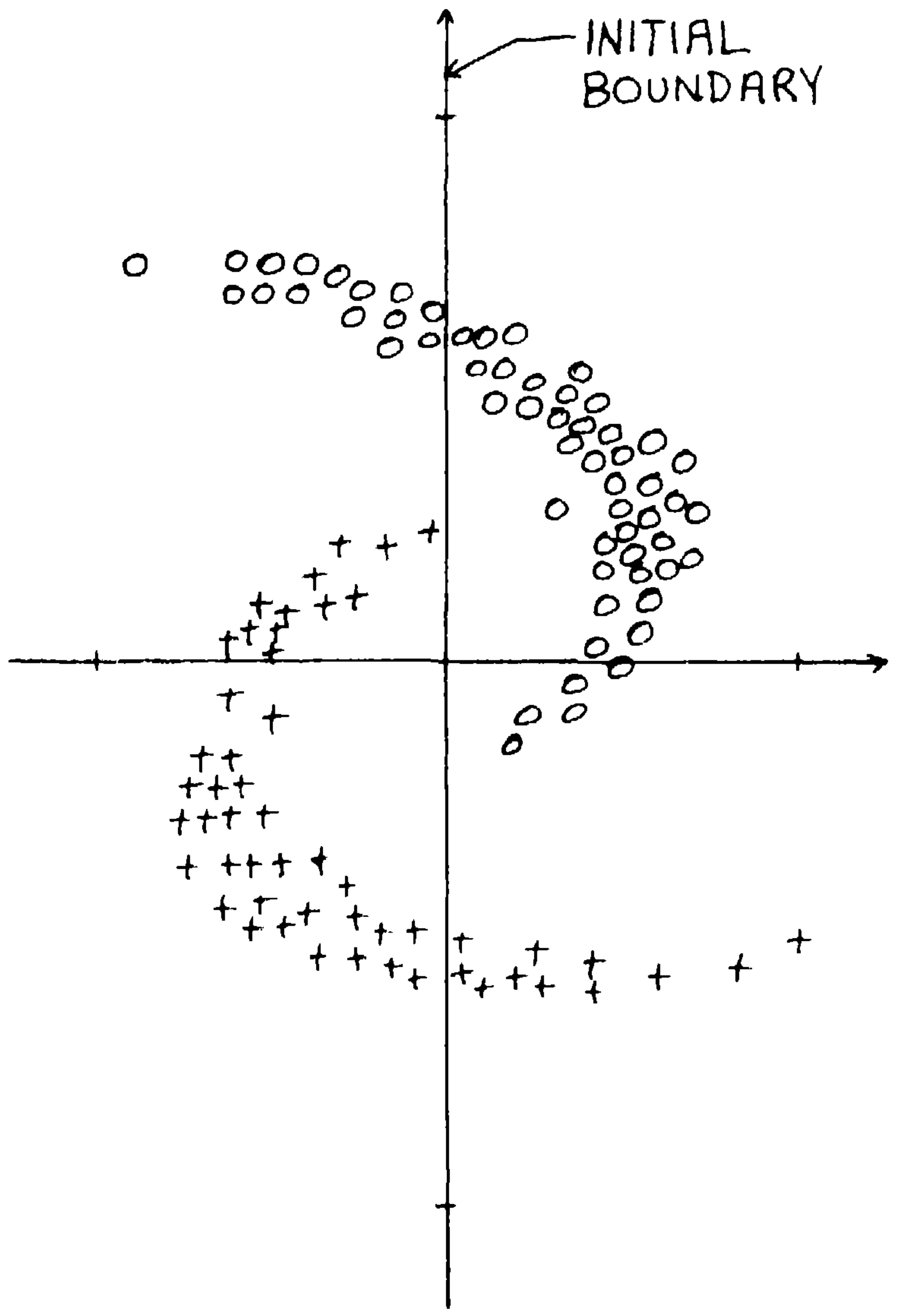

Figure A.
Classification of nonlinearly separable data