September 1994

Patern Recognition Letters 15 (1994) 893–899

How to choose a representative subset from a set of data in multi-dimensional space

B.B. Chaudhuri

Electronics and Communication Sciences Unit, Indian Statistical Institute, 203, B.T. Road, Calcutta 700 035, India

Abstract

Given a set of N points in multi-dimensional space, it may be necessary to choose a subset of n representative points. For example, in clustering problems, it is necessary to choose a few seed points around which the cluster may grow. This problem may be posed as that of choosing one out of each k data when $\lfloor N/n \rfloor = k$. In our proposed method, the data points are ordered in decreasing magnitude of density. The datum toping the ordered list is chosen and its $k-1$ nearest neighbours are deleted from the ordered list. From the remaining data, the one currently toping the list is chosen. The process is repeated till the data are exhausted. The problem of more general choice of n is also addressed.

Keywords: Clustering; Seedpoint selection; Pattern recognition

1. Introduction

Given a set S of N data in $\mathbb{R}^m$ space, we address here the problem of selecting a subset $R \subset S$ of $n < N$ "good representative" data. As compared to N, we may be requested for a small, medium or large n depending on the application requirement. Some examples are as follows.

Small n (Seed point detection). When a few seed points should be selected to initialize a data clustering algorithm.

Medium n (Algorithm economy). A medium value of n may be chosen when N is very large and it is wise to use a costly data processing algorithm on a set of medium size representative data.

Large n (Outlier rejection). When n is large so that $N-n=n'$ is small, the problem may be viewed as that of rejecting n' outliers from the data set.

In some problems, n may be defined in percent of the original data. Also, in some situations the problem is as follows: From a set S of data, how to select (reject) one out of $k \geq 2$ data

Apart from the applications cited above, the problem may be viewed as that of data compression. If we select one out of k data, we can define a cost to be paid for rejecting the $k-1$ data. Thus, the problem is to choose the data for which the cost is minimum. In conventional data compression, say by using the Karhunen–Loeve transform, the number of data remains the same but they are represented in a transformed space of reduced dimensionality and hence less storage space is needed to represent them. In contrast, here the dimensionality is unaltered while the number of data is reduced and hence less storage space is needed to represent them.

It is difficult to define "good representative" points that can satisfy the requirement of a wide variety of applications. A plausible definition is given and some early related work is reviewed in Section 2. Then, in Section 3 our approach is presented.

2. Good representatives and seed points

Let q be a member of the representative data set R. We assume that q represents a subset $S_q \subset S$ of data. The representation is such that for any datum $r \in S_q$, r is closer to q than to any other member of R. In other words, q is the nearest neighbor of the members of S_q.

Since fewer data will represent the space i.e. $R \subset S$, we can consider an error of representation. For the point q, the error is given by

$$e_q = \sum_{r \in S_q} d(q, r) , \tag{1}$$

where $d(q, r)$ is the euclidean distance between q and r given by

$$d(q, r) = \left\{ \sum_{i=1}^{m} [x_i(q) - x_i(r)]^2 \right\}^{1/2} , \tag{2}$$

where $x_i(q)$ is the ith coordinate value of q. The total error is given by

$$E = \sum_{q \in R} e_q . \tag{3}$$

For a given n, the best R is the one for which E is minimum. A straightforward algorithm is not attractive since one has to examine N_{C_n} combinations to get an optimum R. The cost of the algorithm is, therefore, as high as $O(N^{n+1})$ and even for the detection of seed points (where n is small), heuristic approaches should be tried. We briefly review some of the seed point detection techniques.

MacQueen (1967) chooses the first n data units in the data set as the n seed points. Another idea is to label the data as 1, 2, ..., N and choose those labeled N/n, $2N/n$, ..., $(n-1)N/M$ and N. A variation of the labeling approach is to choose the data corresponding to n different random numbers in the range of 1 to N. However, nothing can be said about "good representation" in the above techniques. Astrahan (1970) proposed a better approach where the "density" at each datum is computed as the number of data units within some specified distance, say d_1. The data units are ordered by decreasing magnitude of density and the one with highest density is chosen 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 be at least a minimum distance, say d_2 from all other previously chosen points. The problem with this method lies in the choice of d_1 and d_2 for the stipulated number of seed points. Ball and Hall (1967) suggested a simpler approach where they choose the overall mean vector of the data set as the first seed point. Then any data unit which is at least at some specified distance, say d from all previously chosen seed points, can be chosen as the subsequent seed point. However, it is difficult to say why in this method the chosen data are good representatives except that they span the whole data space.

If the good representatives are defined on the basis of minimization of E in Eq. (3) then we suggest the following approach of obtaining a semi-optimum solution. Select n data at random from the data set S and compute E using Eqs. (1)–(3) K number of times. The set of n data for which E is minimum is finally accepted as the best set of representatives. The computational requirement is much less than that in the optimum method but K should be fairly large to get a good result. We propose another approach in the following section.

3. Proposed approach

Let the problem be that of choosing 1 out of k data units. We propose here a "density" based method with the following steps.

(1) Find the density at each datum and order the data in decreasing magnitude of density. Let L be the ordered list. Let $i \leftarrow 1$.

(2) Choose the datum that tops the list L as the ith representative datum.

(3) Count the number of data in the current S. If the number is less than $k-1$ then stop. Else, from the current S find the $k-1$ nearest neighbours of the datum P which has been chosen in Step 2. Delete P and these $k-1$ neighbours from L and S to get the list of L and S for the next iteration. Make $i \leftarrow i+1$ and go to Step 2.

When we want to reject 1 out of k data, Step 1 should be modified so that L denotes the list in order of increasing magnitude of density. Step 2 and Step 3, however, remain unaltered.

Apparently, there exists some similarity of the above algorithm with that proposed by Astrahan

(1970). In both methods the data are ordered in terms of their density. But in (Astrahan, 1970), the data are chosen so that they are pairwise at least d_2 distance away while in our method the data are chosen by excluding $k-1$ nearest neighbours. The problem of choosing d_2 is avoided here. Thus, if n seed points should be chosen (n is much smaller than N in this case) then our algorithm can be used by considering k as the integer part of N/n. Another advantage of our method is that exactly n seed points can be chosen while no such guarantee is given in the method of Astrahan (1970) even by varying d_2.

If the problem is to select $x\%$ of the data, then too the above algorithm can be used with small modifications. If $x > 50$, then the problem is modified to that of "rejecting" $(100-x)\%$ data. Thus, without loss of generality, we assume that $x \leqslant 50$. If x is such that $x\% = 1/r$ where r is an integer then the above method can be used directly to find the representative points. For example, if $x=25$, then $r=4$ and we can select one out of four data. On the other hand if $x=22$, then we can select one out of r_1 data k_1 times and one out of r_2 data k_2 times so that on an average 22% data are selected. Of the four unknown integers, r_1 and r_2 are chosen as the smallest and largest integers, respectively so that $1/r_1 > x\%$ and $1/r_2 < x\%$. In our example, $r_1=4$ and $r_2=5$ because $\frac{1}{4} > x\%$ and $\frac{1}{5} < x\%$. Now, selecting one out of $r_1=4$ data k_1 times means selecting k_1 out of $4k_1$ data. Then on an average, k_1+k_2 out of $4k_1+5k_2$ data are being selected in our method. According to our requirement

$$\frac{k_1+k_2}{4k_1+5k_2} = 22\% = \frac{22}{100}, \tag{4}$$

which leads to

$$\frac{k_1}{k_2} = \frac{5}{6}. \tag{5}$$

One solution of this relation is to choose $k_1=5$ and $k_2=6$.

Then our algorithm should be modified so that on an average one out of 4 data are chosen 5 times while one out of 5 data are chosen 6 times. Clearly, the modification can be done quite easily.

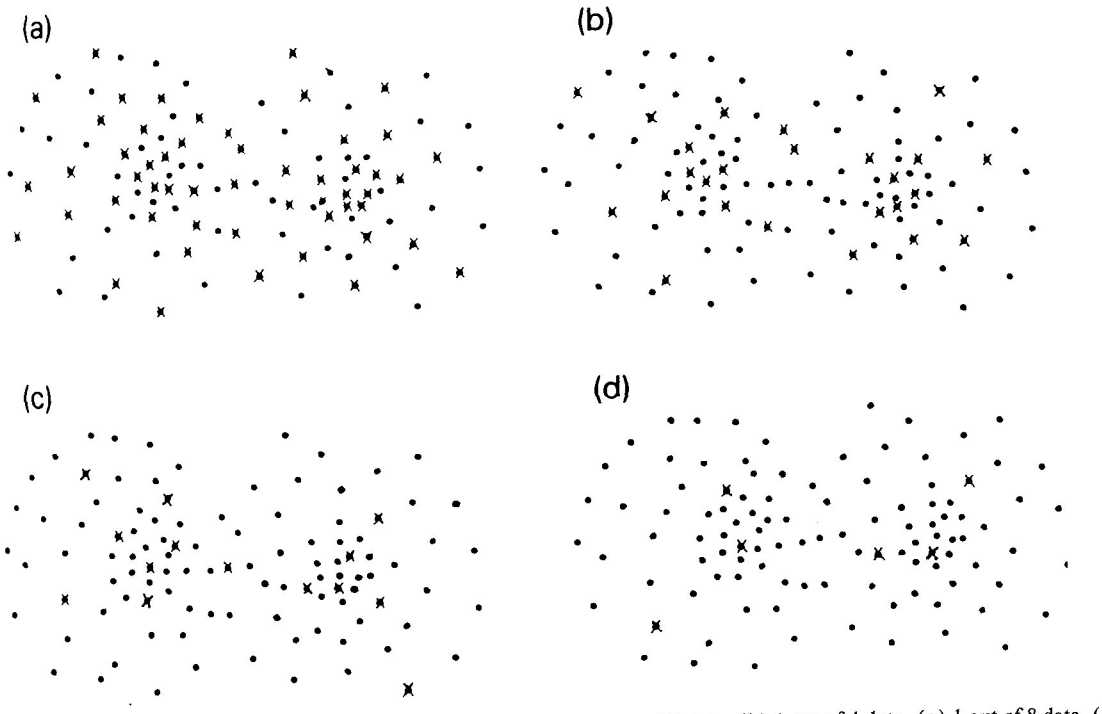

Fig. 1. Representative point selection from touching Gaussian clusters: (a) 1 out of 2 data, (b) 1 out of 4 data, (c) 1 out of 8 data, (d) 1 out of 20 data. (The representative points are shown by ×.)

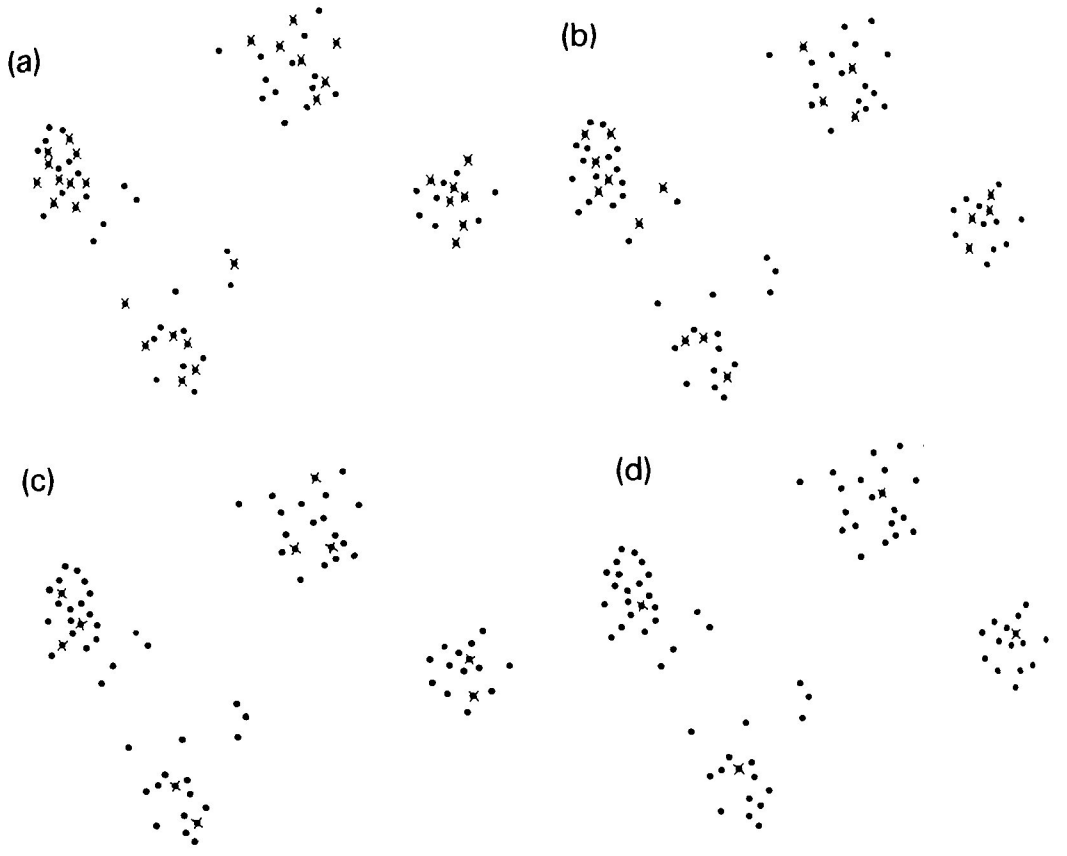

Fig. 2. Representative point selection from another set of data: (a) 1 out of 2 data, (b) 1 out of 4 data, (c) 1 out of 8 data, (d) 1 out o 20 data. (The representative points are shown by ×.)

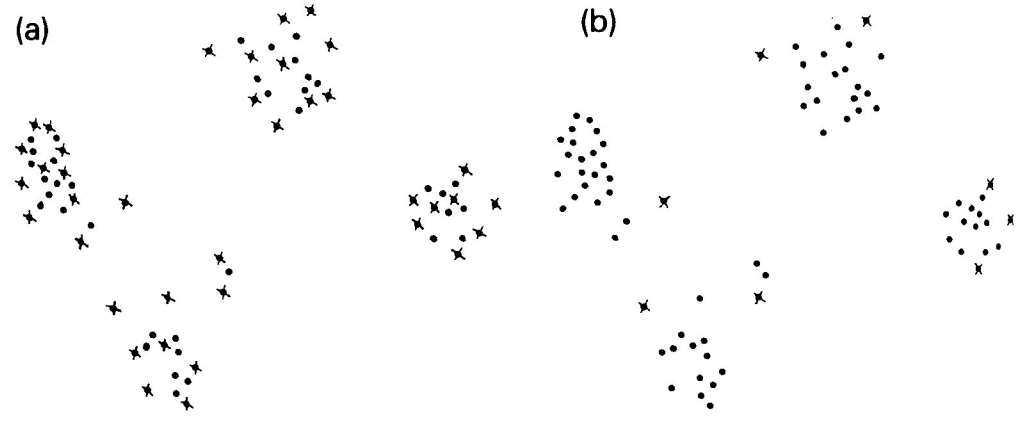

Fig. 3. Rejection of points from the data set of Fig. 2: (a) rejecting 1 out of 2 data, (b) rejecting 1 out of 10 data. (The rejected point are shown by ×.)

4. Results and discussion

The method has been tested on a set of points in space. Two sets of representative results are shown here. Data pooled from a mixture of two Gaussian distributions generate the first set, as shown in Fig. 1. The second set is an arbitrary data denoting clusters of various size. Experiment is also done on rejecting one out of k data, as shown in Fig. 3. We have used the method for seed point detection in connection with data clustering methods similar to the k-means technique and found that they lead to satisfactory results.

Somewhat similar but alternative methods can be proposed for the representative point detection problem. For example, we can choose the data so that their $k-1$ neighbours are mutually "disjoint". However, we may obtain less than N/k data satisfying this condition. Another alternative is to take the first k data from the list (in which the data are ordered according to density), and choose the datum which is nearest to the centroid of these k data. Delete these k data from the list and repeat the process.

A discussion on density estimation is in order. Two popular approaches of density estimation are the Kernal estimator and the k-nearest neighbour estimator (Prakasa Rao, 1983). Although both estimations are consistent and unbiased, there is a difficulty in choosing the window size h in the Kernel estimation or the value of k in the k-nearest neighbour estimator. We used a Kernel-based estimator where the window size h is related to the average edge length of the minimum spanning tree of the data set. In particular, in two dimensions $h=\sqrt{l/N}$, where l is the length

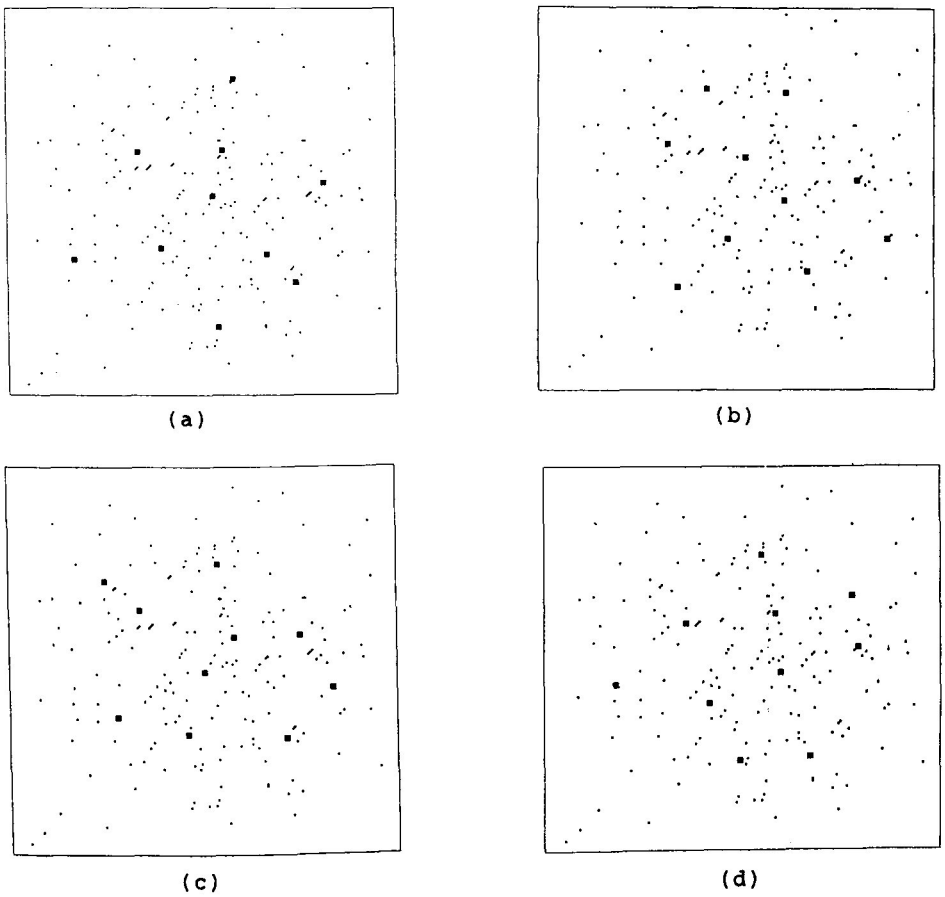

Fig. 4. The effect of changing the Kernel window width in obtaining representative points: (a) window width $2h$, (b) window width h, (c) window width $0.5h$, (d) window width $0.25h$, where h is the window width chosen by the minimal spanning tree method.

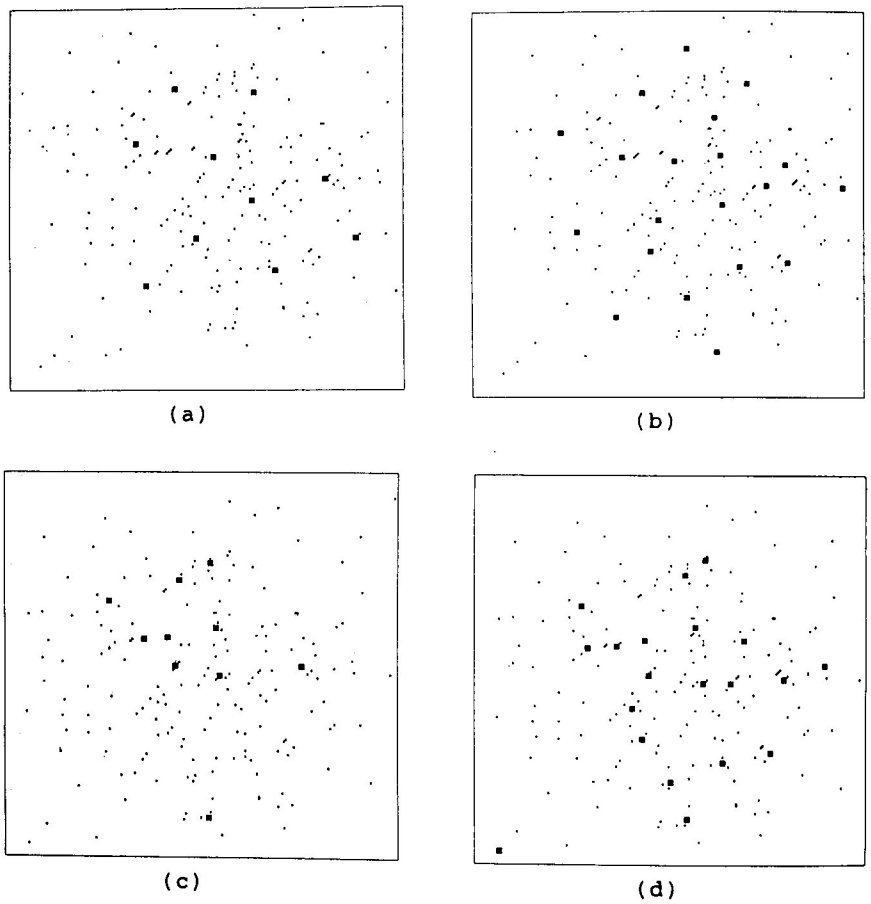

Fig. 5. Comparison of the proposed method with that due to Astrahan: (a) 5% representative, and (b) 10% representative points by the proposed method, (c) 5% representative, and (d) 10% representative points by the method due to Astrahan. (The density estimation in all cases is the same.)

of the minimum spanning tree. In a separate paper (Chaudhuri et al., 1993) we proved that such an estimator is also consistent.

We have tested the effect of varying the window width for density estimation on our method of representative point selection. The results are shown in Fig. 4. Although the chosen representatives are different for different values of window width, they are evenly distributed in each case, which is desirable.

We have compared the method of Astrahan with that of ours where 5% and 10% representatives are to be chosen from 200 random points. The density estimation was done by a Kernel-based method in both cases where the width of the window is $h=\sqrt{l/N}$ where l=length of the minimal spanning tree of the data and N=the number of data points. However, if we use $d=2h$ in Astrahan's method, we can not get 5%, i.e. 10 representative points. In the current example, $d=0.5h$ was chosen for the desired result. Similarly, for 10% representative points, $d=0.25h$ was needed. The results are compared in Fig. 5, where our method shows more homogeneous distribution in representative points. We obtained similar results on other data sets also.

It may be an interesting theoretical study to see if the n data chosen by the proposed approach has the distribution identical to that of the original data as N tends to infinity.

References

Astrahan, M.M. (1970). Speech analysis by clustering, or the hyperphomene method. Stanford A.I. project Mem. Aim-124, Ad 709067. Stanford University, CA.

Ball, G.H. and D.J. Hall (1967). PROMENADE, An one-line Pattern Recognition. Rep. no. RADC-TR-67-310, Ad 822174, Stanford Res. Inst., Menlo Park, CA.

Chaudhuri, D. C.A. Murthy and B.B. Chaudhuri (1993). A minimum spanning tree based probability density estimation procedure. *IEEE Trans. Pattern Anal. Machine Intell.*, communicated.

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. *Proc. 5th Symp. Math., Statistics and Probability*, Vol. 1, Ad 669871, Univ. of Calif. Press, Berkeley, 281–297.

McRae, D.T. (1971). MIKCA: A FORTRAN IV iterative *K*-means cluster analysis program. *Behavioral Sci.* 16 (4), 423–424.