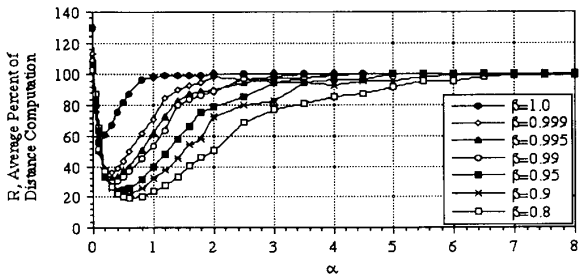

Fig. 4. R, the average percentage of distance computations, as in (7), with a truncated Gaussian kernel function and truncation level 3. The parameter α in the cluster creation condition of (6) is varied from 0.01 to 8. Window size $h = 0.304$.

70% of the distance computations can be saved, as seen in Fig. 4. As seen before with the Epanechnikov kernel function, extremely small or large values of α were not acceptable, because they produced too many small clusters or just a few large clusters. With α in the range of $0.2 \sim 1.0$, it was observed that about $40 \sim 80\%$ savings in distance computation could be achieved.

IV. Conclusion

In this correspondence, a computationally efficient Parzen density estimation algorithm is developed by using a simple branch-and-bound procedure applied to the preclustered data samples. Not only those kernel functions having finite support for nonzero values, such as the Epanechnikov kernel function, but also those kernel functions having nonzero values over the entire feature space, were applicable to this algorithm through truncation. By choosing a proper parameter value for cluster generation, substantial savings in computation could be realized. Values that were found to be satisfactory were those close to the critical distance D_c. Experimental results verified that savings were significant. To further enhance the computational efficiency, this proposed algorithm can be used in conjunction with the data reduction technique [4].

References

[1] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.

[2] P.E. Hart, "The condensed nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 515–516, 1968.

[3] K. Fukunaga and P.M. Narendra, "A branch and bound algorithm for computing K-nearest neighbors," *IEEE Trans. Comput.*, vol. C-24, pp. 750–753, 1975.

[4] K. Fukunaga and R.R. Hayes, "The reduced parzen classifier," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. PAMI-11, pp. 423–425, 1989.

[5] B.W. Silverman, "Kernel density estimation using the fast fourier transform," Statistical Algorithm AS176, *Appl. Statist.*, vol. 31, pp. 93–97, 1982.

[6] M.C. Jones and H.W. Lotwick, "A remark on algorithm AS 176: Kernel density estimation using the fast Fourier transform," Remark, AS R50, *Appl. Statist.*, vol. 33, pp. 120–122, 1984.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.

[8] G.H. Ball and D.J. Hall, "Isodata: A novel method of data analysis and pattern classification," Stanford Research Institute Tech. Rep. NTIS–AD–699616, Stanford, CA, 1965.

[9] C.W. Therrien, *Decision, Estimation, and Classification: An Introduction to Pattern Recognition and Related Topics*. New York: Wiley, 1989.

A Least Biased Fuzzy Clustering Method

Gerardo Beni and Xiaomin Liu

***Abstract*—A new operational definition of cluster is proposed, and a fuzzy clustering algorithm with minimal biases is formulated by making use of the Maximum Entropy Principle to maximize the entropy of the centroids with respect to the data points (*clustering* entropy). We make no assumptions on the number of clusters or their initial positions. For each value of an adimensional scale parameter β, the clustering algorithm makes each data point iterate towards one of the cluster's centroids, so that both hard and fuzzy partitions are obtained. Since the clustering algorithm can make a multiscale analysis of the given data set we can obtain both hierarchy and partitioning type clustering. The relative stability with respect to β of each cluster structure is defined as the measurement of cluster validity. We determine the specific value of β which corresponds to the optimal positions of cluster centroids by minimizing the entropy of the data points with respect to the centroids (*clustered* entropy). Examples are given to show how this least-biased method succeeds in getting perceptually correct clustering results.**

***Index Terms*—Clustering, fuzzy clustering, maximum entropy principle, cluster validity.**

I. Introduction

Cluster analysis attempts to discover an inherent structure in a set of data points as a partition in (or hierarchy of) subgroups of points (clusters). Many algorithms have been proposed for solving this problem [1], and their merits are still debated. The problem is intrinsically difficult because no *a priori* information on the data distribution can be assumed. In most traditional clustering algorithms, the relationship between data points and clusters is represented in one of two ways: a) each data point may belong to one and only one cluster (hard clustering), as e.g., in [2]; b) each data point may belong to one or more clusters with a certain degree of membership (fuzzy clustering), as e.g., in [3]. The latter type of representation is more general and more precise in principle, although its practical value is still not clearly proven. In this correspondence, we will choose the fuzzy clustering approach and obtain a unique hard clustering partition as a byproduct.

Any clustering algorithm (fuzzy or hard) runs typically as follows: *a membership function* generates (usually *assuming* the initial positions of the cluster centroids) a partition; the fitness of this partition is measured by a *cost function* (usually defined in term of the membership function). An iteration scheme is implemented until the membership function generates a partition which minimizes the cost function.

Unfortunately, if the cost function used is not convex and has local minima (a typical case), the algorithm may be trapped into one of them, resulting in a nonoptimal partition. Obtaining an optimal partition with the cost function converging to the global minimum, depends on whether or not a "right" number of cluster centroid for the centroids has been assumed at the beginning of the algorithm; and it depends also on whether or not these centroids have been positioned properly. Our algorithm is independent of such initial choices.

Manuscript received November 4, 1992; revised June 5, 1993. Recommended for acceptance by Editor-in-Chief A. K. Jain.

The authors are with the College of Engineering, University of California, Riverside, Riverside, CA 92521-0425 USA; e-mail: Gerardo_Beni@qmail.ucr.edu.

IEEE Log Number 9401639.

Recently, some efforts aimed at this nonconvex optimization problem have been made successfully. The simulated annealing method introduced in [4] has in principle the capability of achieving a global minimum [5], if the schedule obeys $T \propto 1/\log n$ (where T is the temperature and n is the number of the current iteration). Such schedules are not realistic in many applications. A more practical algorithm, the deterministic annealing approach, has been proposed in [6]. Both algorithms [4], [6] operate under the control of a scale parameter analogous to the temperature in statistical mechanics. The capability of the deterministic annealing method [6] to avoiding local minima is higher than that of other popular techniques, e.g., the K-means algorithm [7], but the minimization of its cost function still may encounter local minima traps.

It is important to note that in the traditional clustering algorithms, even though *a priori* knowledge may not play a role, the user's judgment plays a critical role. In fact, the number of cluster centroids as well as their initial positions are assumptions made by the user. With such assumptions, the algorithm discovers a cluster structure within the data set, but the chosen cluster number may be incorrect, or the data set may be totally at random and hence "unclusterable." Thus, the answers to two basic questions "How many clusters?" and "Are the data at random?" are generally provided by the user and not by the algorithm. This is not the case for the algorithm presented here.

To evaluate the user's assumptions, typically, clustering tendency and cluster validity are tested [8], [9]. Most formulations of cluster validity, such as those discussed in [8] and [9], measure the ratio of "separation" to "compactness" of a partition. The 'best' partition is the one which optimizes such a validity function. The difficulty is that a validity function measuring objectively the intuitive notion of separation/compactness is difficult to define. In this paper the criterion for cluster validity (discussed in detail in Sections II-D and E) is inherent in the clustering algorithm itself and does not require any further assumptions.

Generally, the performance of traditional clustering algorithms is also affected by the metric chosen. Although it is well known that different distance measures can lead to different partitions, the distance is typically chosen to be Euclidean [11] and the scale is chosen intuitively from inspection. By choosing a metric in this way, a fundamental subjective bias may be introduced. This issue has been addressed recently [12] where clustering is attempted without a metric. In our algorithm, the metric is not avoided, but it is derived: both scale and form of the distance function are calculated as intrinsic properties of the data set.

Fuzzy clustering techniques may also induce biases in the choice of the membership function. The Maximum Entropy Principle (MEP) [13] which ensures a maximum of fairness in handling "lack of information," is used in this paper to minimize such bias in the choice of the membership function, as shown in Section II-B.

In this correspondence, we propose a clustering algorithm based on a *resolution parameter* β. For a given resolution β, a subgroup of data points (a cluster) converges towards one fixed point (their cluster's centroids). The relative stability of a partition with respect to β is used to measure its validity. Since each data point associates each cluster with a grade of membership, our algorithm may be regarded as a type of fuzzy clustering based on a new operational definition of cluster (Section II-A).

The rest of the correspondence is organized as follows. In Section II, we discuss the formulation of the clustering algorithm. After the new definition of cluster, we introduce the cluster membership via the Maximum Entropy Principle (Section II-B); the bounds on resolution parameter β are discussed next (Section II-C). The cluster validity measure is presented in Sections II-D and E and the overall algorithm in Section II-F. In Section III, we give some representative examples of the algorithm and compare the clustering results of this algorithm with the fuzzy c-means clustering algorithm of [3]. In the conclusion (Section IV) we summarize the method and the main findings.

II. Least Biased Fuzzy Clustering Method

A. No-bias Operational Definition of Cluster

In fuzzy clustering methods, typically the N points of a data set $\{\vec{x}_i; i = 1, 2 \cdots N\}$, are related to the γ cluster centroids $\{\vec{c}_m, m = 1 \cdots \gamma\}$ via membership functions. It is convenient to visualize each cluster centroid $\vec{c}_m$ as "active" in trying "to own" (i.e., "to cluster") data points by distributing its total (normalized to 1) membership among them, which we call the *clustering* membership $p_{mi}(\vec{x}_i, \vec{c}_m)$ (i.e., the probability that centroid m will cluster the data point i). As a result, each data point is "owned" (i.e., "clustered") by every centroid to an extent given by what we call the *clustered* membership μ_{im}. Numerically, clustering and clustered memberships are identical (i.e., $p_{mi}(\vec{x}_i, \vec{c}_m) = \mu_{im}$) but the latter is not normalized.

We now make the plausible assumption (the only one of the method) that the centroids have *no-bias* towards any of the points; hence they position themselves at zero average distance from all the data points, i.e., the possible locations of the centroids satisfy the condition:

$$\langle \vec{x}_i - \vec{c} \rangle = 0 \tag{1}$$

where $\vec{c}$ is the location of the centroid and the average $\langle\ \rangle$ is calculated over the clustering membership distribution for all N data points. Equation (1) is our no-bias operational definition of cluster centroid. The entire clustering algorithm, including the metric and the number of centroids, follows from (1).

Since we do not specify the number of clusters in the data set, $\vec{c}$ will be used to represent the centroid of any possible cluster for the time being. Hence, we denote the clustering membership $p_{mi}(\vec{x}_i, \vec{c}_m)$ by $p_i(\vec{x}_i, \vec{c})$, which satisfies the normalization condition:

$$\sum_{i=1}^{N} p_i(\vec{x}_i, \vec{c}) = 1. \tag{2}$$

and rewrite (1) as

$$\sum_{i=1}^{N} (\vec{x}_i - \vec{c}) p_i(\vec{x}_i, \vec{c}) = 0. \tag{1a}$$

Since we do not wish to make any further assumptions besides that of no-bias, we seek a no-bias form of the clustering membership. For this we shall use the minimally prejudiced (least presumptive) probability distribution for describing the clustering membership. Thus, we apply the Maximum Entropy Principle (MEP) [13] under the constraints (1a) and (2). The entropy of the clustering membership distribution (which we shall call the *clustering entropy*) of a centroid at location $\vec{c}$ is

$$S = -\sum_{i=1}^{N} p_i(\vec{x}_i, \vec{c}) \log p_i(\vec{x}_i, \vec{c}). \tag{3}$$

In the next subsection we show how the centroid numbers, location, clustering memberships and "hard" clustering partitions can be derived from (1a), (2), and (3) by using the MEP, without any further assumptions.

B. Membership Distribution from MEP

The constraint (1a) is a vector equation. If the spatial dimension is K, then (1a) will be equivalent to K equations:

$$\sum_{i=1}^{N}(x_{i\alpha}-c_\alpha)p_i(\vec{x}_i,\vec{c})=0. \qquad \alpha=1\cdots K. \tag{4}$$

Further, the K equations in (4) are equivalent to 2K equations:

$$\sum_{i=1}^{N} f_\alpha^+(x_{i\alpha},c_\alpha)p_i(\vec{x}_i,\vec{c})=A_\alpha, \qquad \alpha=1\cdots K; \tag{5a}$$

$$\sum_{i=1}^{N} f_\alpha^-(x_{i\alpha},y_\alpha)p_i(\vec{x}_i,\vec{c})=-A_\alpha, \qquad \alpha=1\cdots K. \tag{5b}$$

where A_α is an arbitrary constant $A_\alpha \geq 0$ and the *distance functions* f_α^+ are defined as follows:

$$f_\alpha^+(x_{i\alpha},c_\alpha)=\begin{cases} x_{i\alpha}-c_\alpha, & x_{i\alpha}-c_\alpha>0. \\ 0. & x_{i\alpha}-c_\alpha<0. \end{cases} \tag{6a}$$

$$f_\alpha^-(x_{i\alpha},c_\alpha)=\begin{cases} 0. & x_{i\alpha}-c_\alpha>0. \\ x_{i\alpha}-c_\alpha, & x_{i\alpha}-c_\alpha<0. \end{cases} \tag{6b}$$

Equations (6a), (6b) are used to separate the data points into two groups for each spatial dimension α; one (the other) group is for the points whose values $x_{i\alpha}$ are less (more) than c_α, i.e., at the "left" ("right") of $\vec{c}$. Note that we do not make any restriction on how to choose the value of A_α for each spatial dimension α. Equations (5a), (5b) will be used as our constraint functions instead of (1a).

We can now determine the form and magnitude of the clustering membership distribution $p_i(\vec{x}_i,\vec{c})$ by applying the MEP to the clustering entropy S, (3), under the $2K+1$ constraints of (2), (5a), (5b). It is well known that entropy maximization yields a distribution of the Gibbs type; in fact by maximizing the clustering entropy we obtain:

$$p_i(\vec{x}_i,\vec{c})=Z^{-1}\exp\sum_{\alpha=1}^{K}[-\beta_\alpha^+ f_\alpha^+(x_{i\alpha},c_\alpha) + \beta_\alpha^- f_\alpha^-(x_{i\alpha},c_\alpha)]. \tag{7}$$

where $\beta_\alpha^\pm$ are $2K$ Lagrange multipliers for $\alpha=1\cdots K$; and Z is the partition function:

$$Z=\sum_{i=1}^{N}\exp\sum_{\alpha=1}^{K}[-\beta_\alpha^+ f_\alpha^+(x_{i\alpha},c_\alpha)+\beta_\alpha^- f_\alpha^-(x_{i\alpha},c_\alpha)]. \tag{8}$$

Given a set of A_α for $\alpha=1\cdots K$, the Lagrange multipliers $\beta_\alpha^\pm$, which "physically" represent *resolution parameters*, can be determined by the equations:

$$-\frac{\partial \log Z}{\partial \beta_\alpha^\pm}=A_\alpha, \qquad \alpha=1\cdots K. \tag{9}$$

Thus, assigning the K parameters A_α is equivalent to assigning the 2K resolution parameters $\beta_\alpha^\pm$. In fact, for the spatial dimension α, β_α^+ is the resolution applied to the group of data points at a "positive distance" from the centroid. Similarly, β_α^- is the resolution applied to the group of data points at a "negative distance" from the centroid. Since there is no reason to expect any spatial anisotropy between the directions right and left of the centroid (i.e., there is no left or right bias), it is expected that indeed $\beta_\alpha^+=\beta_\alpha^-=\beta_\alpha$ for $\alpha=1\cdots K$.

Thus we obtain a *vector resolution parameter* $\vec{\beta}$ on the data space. By manipulating this resolution vector, we can deal with the difficulties raised by inhomogeneous scales in feature space. This will be discussed in detail in another paper. In this correspondence, we restrict our applications to a homogeneous feature space; thus, the resolution vector $\vec{\beta}$ is reduced to a scalar by $\beta_\alpha=\beta$, for $\alpha=1\cdots K$. In this way, our equations in (7) and (8) are simplified to:

$$p_i(\vec{x}_i,\vec{c})=Z^{-1}\exp[-\beta D(\vec{x}_i,\vec{c})], \tag{7a}$$

$$Z=\sum_{i=1}^{N}\exp[-\beta D(\vec{x}_i,\vec{c})], \tag{8a}$$

where β is the *resolution parameter*, and D is city-block distance between $\vec{x}_i$ and $\vec{c}$ D is defined as

$$D(\vec{x}_i,\vec{x}_j)=\sum_{\alpha=1}^{K}|x_{i\alpha}-x_{j\alpha}|. \tag{10}$$

Given β and the position of a centroid $\vec{c}$, the cluster membership value $p_i(\vec{x}_i,\vec{c})$ for data point $\vec{x}_i$ in this cluster can be calculated by (7a), (8a). On the other hand, knowing the cluster membership value $p_i(\vec{x}_i,\vec{c})$ of each data point, the location of this cluster's centroid can be determined from the basic constraint equation (1a). By subsitituting the explicit form of $p_i(\vec{x}_i,\vec{c})$ given by (7a), (8a) into (1a) we have

$$\sum_{i=1}^{N}(\vec{x}_i-\vec{c})\frac{\exp[-\beta D(\vec{x}_i,\vec{c})]}{\sum_{i=1}^{N}\exp[-\beta D(\vec{x}_i,\vec{c})]}=0. \tag{11}$$

or

$$\vec{c}=\sum_{i=1}^{N}\vec{x}_i p_i(\vec{x}_i,\vec{c})=\frac{\sum_{i=1}^{N}\vec{x}_i\exp[-\beta D(\vec{x}_i,\vec{c})]}{\sum_{i=1}^{N}\exp[-\beta D(\vec{x}_i,\vec{c})]}. \tag{12}$$

Equation (12) is an implicit nonlinear equation for $\vec{c}$. It leads an initial value of $\vec{c}$ to iterate towards a fixed point:

$$\vec{c}^{(n+1)}=\sum_{i=1}^{N}\vec{x}_i p_i(\vec{x}_i,\vec{c}^{(n)}), \qquad n=0\cdots\infty. \tag{13}$$

Convergence to the same fixed point is not guaranteed by starting from an arbitrary location. However, we have found that, by choosing the initial location to coincide with a data point, the convergence is unambiguous for a given resolution. The algorithm to find how many centroids satisfy (12) for a specific resolution β is the following.

1) Find the position of the centroid $\vec{c}$ by iterating (12) starting from $\vec{c}^{(0)}=\vec{x}_i$ where $\vec{x}_i$ is any data point.
2) Repeat step 1 for all data points.
3) Classify the results. The data points starting from which we obtain the same centroid position $\vec{c}$ form the cluster (hard) of that centroid.

In the next section, we will discuss the bounds on the resolution parameter β.

C. Bounds on the Resolution Parameter β

As discussed in previous sections, the resolution parameter β is, so far, chosen arbitrarily since there is no bias on the choice of the constant A_α in the constraint equations (5a), (5b). However, limits on the maximum value of β can be imposed. For example, in every physical measurement the measured values are only known to within the limits of the experimental uncertainty. Thus, the resolution β, in order to be meaningful, should not exceed $2/\varepsilon_{\min}$, where $\varepsilon_{\min}$ is the accuracy of measurement of the data set. If we denote this value of the resolution as $\beta_{\max}$, then, for $\beta>\beta_{\max}$, the resolution can be considered *redundant* and hence discarded.

Besides by the experimental uncertainty, the maximum resolution may be limited in other ways. Three notable examples are now given:

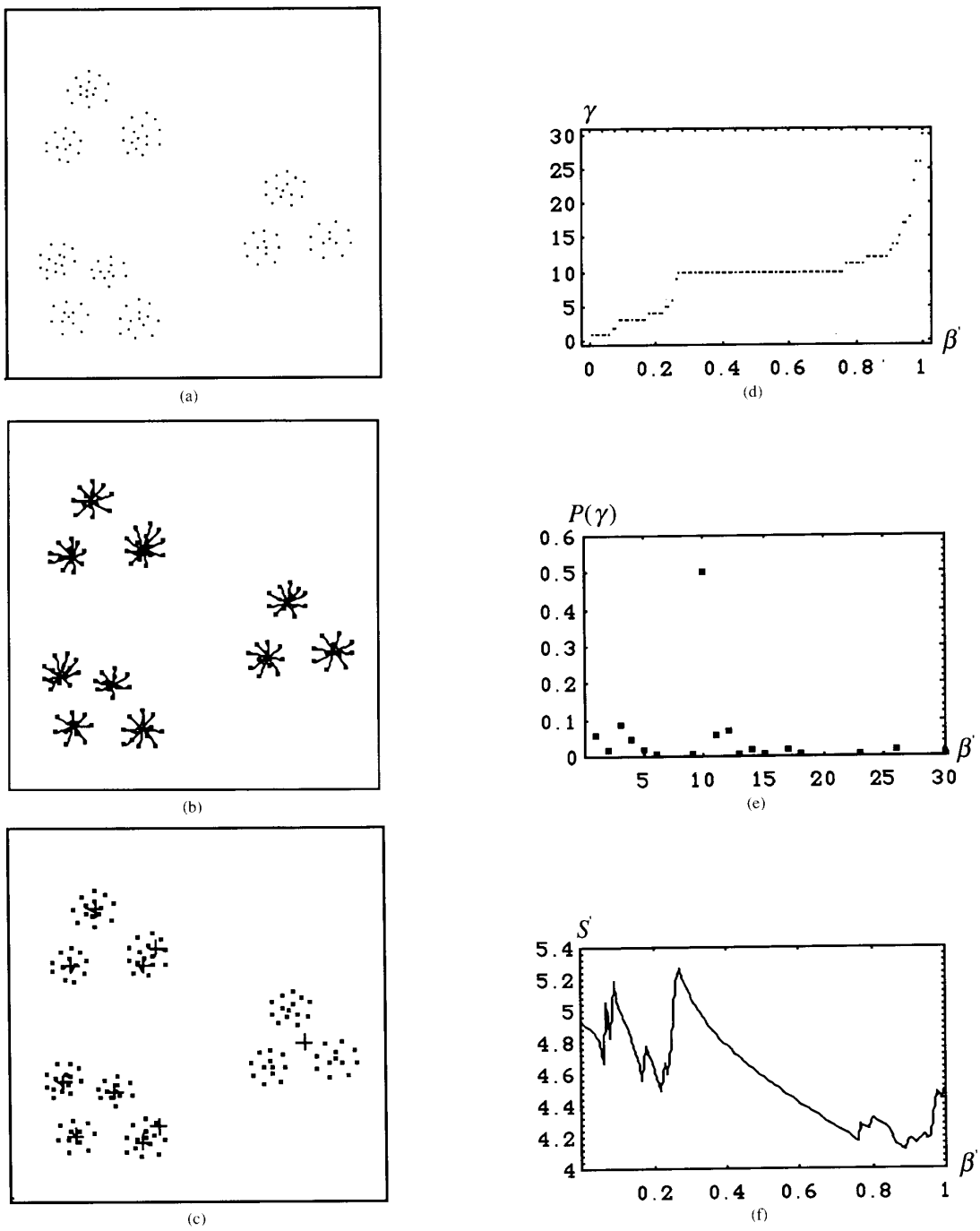

Fig. 1. Clustering results, according to the least biased clustering algorithm, as applied to the data of (a). (b) Optimal clustering result. (c) The typical partitioning results by the fuzzy c-means clustering algorithm in [3]. (d) Number of clusters as a function of the resolution parameter for the data set. Note that the partition remain with ten clusters for the longest range β (from $\beta'_{\min}(10) = 0.27$ to $\beta'_{\max}(10) = 0.76$. (e) Validity diagram. We can see from this diagram that the cluster structure with 10 clusters is the most probable one. (f) Clustered entropy showing minimum at $\beta'(10) = 0.76$.

1) the experimental resolution, i.e., $\beta_{\max} = 2/\varepsilon_{\min}$;
2) the data resolution, i.e., $\beta_{\max} = 2/D_{\min}$, where $D_{\min}$ is the minimum distance between points;
3) the most cost-effective resolution, i.e., $\beta_{\max} = 2^{\lambda'+1}/D_{\max}$, where $D_{\max}$ is equal to the largest range for all the feature dimensions such that all the data points are included into a hypercubic box of side $D_{\max}$; if the box is subdivided in a binary fashion, and N_β is the number of byperbins with one or more points; then among all the number of binary divisions λ, λ' is the one yielding the maximum of $\Delta N_\beta(\lambda)/\Delta\lambda$.

The results of the examples presented in Section III are based on choice 3 for $\beta_{\max}$.

In any case, the aspects of the clustering algorithm discussed in the rest of part 2 are independent of the choice of $\beta_{\max}$.

We define $\beta' = \beta/\beta_{\max}$ and consider β' in the range $0 \leq \beta' \leq 1$.

By substituting β' in the basic clustering formula (12) we can write:

$$\vec{c} = \sum_{i=1}^{N} \vec{x}_i p_i(\vec{x}_i, \vec{c}) = \frac{\sum_{i=1}^{N} \vec{x}_i \exp[-\beta' D'(\vec{x}_i, \vec{c})]}{\sum_{i=1}^{N} \exp[-\beta' D'(\vec{x}_i, \vec{c})]}. \quad (14)$$

where D' is an adimensional city-block distance given by

$$D'(\vec{x}_i, \vec{x}_j) = \sum_{\alpha=1}^{K} \beta_{\max} |x_{i\alpha} - x_{j\alpha}|. \quad (15)$$

We will use (14) instead of (12) as the form of the basic equation in the rest of the description of the clustering algorithm.

D. Validity of the Number of Clusters

Since we are making no assumptions on the value of the resolution parameter we solve (14) for *all* the values of β' in the range $0 \leq \beta' \leq 1$; *all* is meant to within the chosen, finite numerical quantization of the values of β. Let $1/Q$ be the quantization parameter so that the total number of considered values of β, and of solutions to (14), is Q (Note that we have chosen $Q = 100$ for all the examples presented in this correspondence).

We regard these Q solutions as equally likely events. All thesolutions yielding the same number of clusters γ can be counted: $p(\gamma)$. Their fraction $P(\gamma)$ of the total Q. $P(\gamma) = p(\gamma)/Q$ can be regarded as the probability that the solution to (14) will yield γ clusters when nothing is known about the value of β, which is what we assume since this is a no-bias clustering algorithm. Equivalently, $P(\gamma)$ can be regarded as the measure of the "validity" of γ. Hence, the value of γ corresponding to the maximum of $P(\gamma)$, γ_0, is the most likely number of clusters. Figures 1(d) and (e) illustrate the meaning of $P(\gamma)$. Both figures refer to the data set shown in Fig. 1(a). The configuration can intuitively be viewed as ten clusters. Fig. 1(d) shows that, by increasing β':

1) the solutions yield an increasing number of cluster (as expected); and more importantly:
2) the number of solutions corresponding to a given number of clusters is not (necessarily) increasing. In fact, typically, there is a maximum (and local maxima). This is illustrated in Fig. 1(e) which shows $P(\gamma)$ for the same example. There is a maximum at $\gamma_0 = 10$, which agrees with human perception of the data.

So far, we have established that the cluster partitions with the highest validity are those with γ_0 clusters, i.e., those with the number of clusters most likely to occur. We will establish which of the partitions with γ_0 clusters is the most valid in the next section.

E. Validity of Partitions

In determining the most likely number of cluster, we have looked for partitions that maximize the *clustering entropy*, (3). this corresponds to positioning the clustering centroids without bias with respect to the data points. All the solutions found from (14), including those which we are considering now, i.e., those with γ_0 clusters, have the property. This maximum entropy property corresponds to the tendency to create uniform (e.g., disordered) clusters and it is easy to interpret in physical, e.g., in statistical mechanics models.

On the other hand, the clustering problem is not just a physical problem, but an information problem. Clustering has a meaning

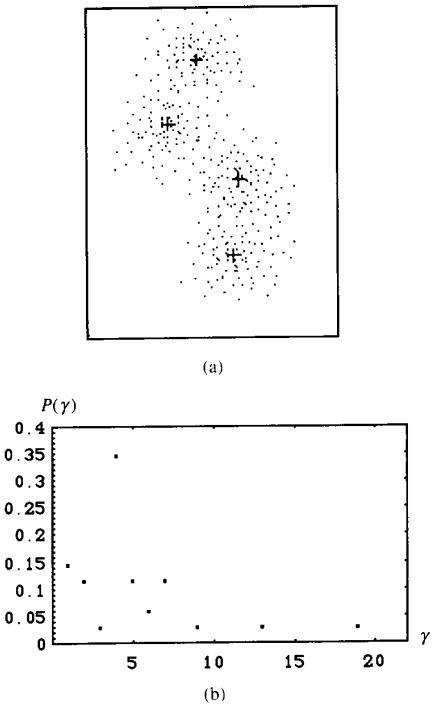

Fig. 2. (a) Four Gaussian clusters with overlapping boundaries. (b) The cluster validity diagram for the data set.

only for an intelligent observer. The intuitive notion of separation of clusters corresponds to a tendency to order (i.e., to increasing information) which is not conveniently modeled mechanically, since it corresponds to decreased entropy. Within our no-bias context, however, it can be quantified as follows. If we regard the data points as having a tendency to be clustered preferentially by only one centroid they will tend to separate. The maximum separation is obtained by *minimizing* the *clustered entropy*, which we define next.

From the definitions of clustering and clustered memberships of Section II-A, we have

$$\sum_{i=1}^{N} \sum_{m=1}^{\gamma_0} \mu_{im} = \sum_{m=1}^{\gamma_0} \sum_{i=1}^{N} p_{mi} = \gamma_0. \quad (16)$$

where γ_0 is the number of clusters found for a specific value of β'. From (16), we have

$$\sum_{i=1}^{N} \sum_{m=1}^{\gamma_0} \frac{\mu_{im}}{\gamma_0} = 1. \quad (17)$$

Thus, we can regard μ_{im}/γ_0 as a probability: the probability for the ith data point to be clustered by the mth cluster centroid. We can define the *clustered* entropy S' for all the data points as

$$S' = -\sum_{i=1}^{N} \sum_{m=1}^{\gamma_0} \frac{\mu_{im}(\vec{x}_i, \vec{c}_m)}{\gamma_0} \log \frac{\mu_{im}(\vec{x}_i, \vec{c}_m)}{\gamma_0}. \quad (18)$$

which can be simplified as follows:

$$S' = -\frac{1}{\gamma_0} \sum_{i=1}^{N} \sum_{m=1}^{\gamma_0} \mu_{im}(\vec{x}_i, \vec{c}_m)[\log \mu_{im}(\vec{x}_i, \vec{c}_m) - \log \gamma_0]. \quad (19)$$

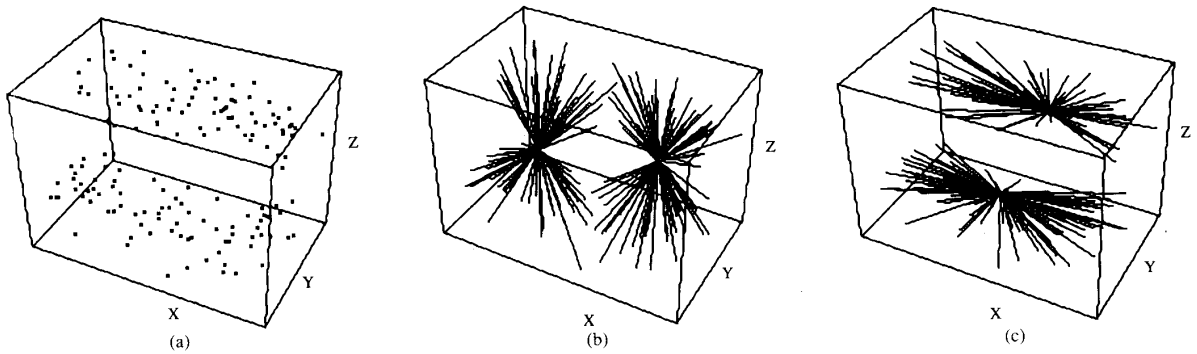

Fig. 3. (a) Three-dimensional "cigar-shaped" data. In the Y and Z directions, the data are distributed as Gaussians. The data points are randomly distributed in the X direction. (b)The partitioning result of the fuzzy c-means algorithm for the data set in Fig. 1(a). The cigar-shaped structure is not recovered. (c) The partitioning result of the least biased fuzzy clustering algorithm. A cigar-shaped clustering result is obtained.

From this, we obtain by noting that $p_{mi}(\vec{x}_i, \vec{c}_m) = \mu_{im}$ (see Section II-A):

$$S' = \frac{1}{\gamma_0} \sum_{m=1}^{\gamma_0} S_m + \log \gamma_0. \tag{20}$$

where S_m is the mth cluster's clustering entropy.

Within the range of values of β' yielding γ_0 clusters we can find the β' corresponding to the minimum value of S' by using (20). However, it turns out that the minimum value of S' always corresponds to the largest value of β' (for $\gamma = \gamma_0$). Hence it is easy to determine this value $\beta'_{\max}(\gamma_0)$ from the validity diagram. This $\beta'_{\max}(\gamma_0)$ corresponds to the most valid partition for $\gamma = \gamma_0$. Figure 1(f) shows that the minimum of the clustered entropy coincides with the largest value of β' for a fixed number of clusters. We can see that the optimal partition is obtained for $\beta'_{\max}(\gamma_0 = 10) = 0.76$ from Fig. 1(b), which shows the optimal (hard) partition.

F. Outline of the Algorithm

For clarity, we recapitulate the main steps of the clustering algorithm.

1) Choose the maximum resolution $\beta_{\max}$. See Section II-C, case 3).
2) For $\beta'(= \beta/\beta_{\max})$ within $0 \leq \beta' \leq 1$, find the number of distinct $\vec{c}$, which satisfies (14). See Sect. II-B.
3) Determine the optimum number of clusters γ_0 from the validity diagram. See Section II-D.
4) For γ_0 clusters, find the maximum value of $\beta' = \beta'_{\max}(\gamma_0)$. See Section II-E.
5) The centroid positions for $\beta'_{\max}(\gamma_0)$ and the associated data points (step 3, Section II-B) form the optimal (hard) clustering partition. The fuzzy clustering partition is obtained from the membership function, (7a) of section II-B.

III. Examples

We have generated additional data sets to test the performance of ourclustering algorithm. The first artificial data set is closer to real world data. It has four clusters with Gaussian distributions as shown in Fig. 2(a). The four clusters are partially overlapping. The validity diagram of the data set from our clustering algorithm in Fig. 2(b) indicates that the most valid clustering result of the data set consists of four clusters. The final positions of the four cluster's centroid calculated by our algorithm are marked by crosses in Fig. 2(a).

The ability of our algorithm to find the optimal position of a cluster's centroid is shown also in the following example. Two clusters with "cigar" shaped data are shown in the Fig. 3(a). The distributions in Y and Z directions are Gaussian. In the X direction, the points are randomly distributed over a limited range. In this case, the fuzzy c-means algorithm [3] cannot find the "correct" centroid's positions of these two clusters. This results is shown by Fig. 3(b), which is obtained after assigning each data point to its nearest centroid. The partitioning result of the least biased clustering algorithm is shown in Fig. 3(c). A cigar-shaped partitioning results is obtained.

The fuzzy c-means clustering algorithm has also been applied to the same data set in Fig. 1(a) (from our algorithm, the optimal partition with ten clusters is obtained as shown in Fig. 1(b). The initial positions of ten centroids were chosen randomly. A typical clustering result is shown in the Fig. 1(c), where the crosses indicate the calculated positions of the ten cluster centroids. Obviously, it is not an optimal clustering result. If the initial positions of the ten centroids were fortuitously chosen to be those obtained from our algorithm, then the fuzzy c-means algorithm would give the "right" clustering result, but the algorithm by itself cannot guarantee this.

We have also applied our algorithm to Fisher's "iris" data [15].The data set consists of two sepal and two petal measurements from 150 irises, 50 from each species (1, Setosa, 2, Versicolor, 3, Virginica). From [15], we know that the group 1 is well separated from groups 2 and 3 but 2 and 3 are overlapped. The cluster validity diagram for this data set is shown in Fig. 4. The diagram indicates that the cluster structure with two cluster is best, the next best one has three clusters, and the next has one. For the cluster structure with two clusters, the first cluster has all the 50 points from Setosa, the second cluster includes 100 data points from Versicolor and Virginica. For the partitioning with three clusters, our algorithm has correctly reclassified 124 out of 150, or 83% of the irises. This result is not as good as the clustering results in [12], which claims 88% correctness. But this comparison cannot be used to judge which algorithm is actually superior, since the judgment is based on some additional information (i.e., the knowledge of the kinds of trees) which is not in the original data set. Without this extra information, one cannot calculate the actual overlapping part of the two clusters.

Finally, the computation load of the algorithm presented here is rather heavy. For each β' of the Q intervals on the entire valid adimensional resolution range, the computation time is of the same order of the fuzzy c-means algorithm, but there are Q such computations to be carried out. A much more efficient computational method for the same algorithm presented in this paper will be discussed in a forthcoming paper [16]. In this method, the cluster number is recovered by selecting some initial centroid's positions

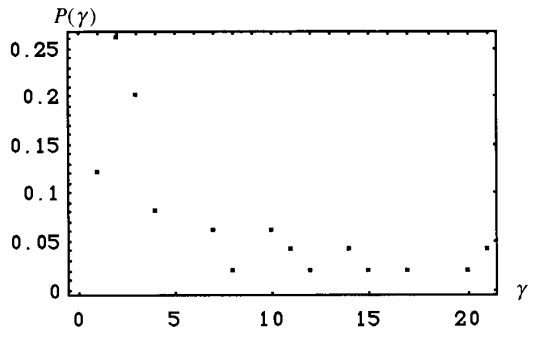

Fig. 4. The cluster validity diagram for the iris data.

rather than using all the data point's positions. The selection of the initial centroid's positions is based on the following heuristics. If the initial centroid position (i.e., the position of a data point) converges (by using the iteration formula in (14)) to a cluster centroid, then the data points which are closer than that point to the cluster centroid will also converge to the same centroid.

IV. Conclusion

Many algorithms have been devised for clustering; all necessarily includingconsiderable subjective bias. In this paper we have presented an algorithm which we believe uses the least amount of subjective bias. The main idea is that the subjective bias is minimized if "ignorance" is treated most fairly. The fairest way of treating ignorance is to optimize the appropriate entropy, as shown by the Maximum Entropy Principle. In our case, we have applied this idea to optimize both the "compactness" and the "separation" of the clustering, in one case maximizing, in the other minimizing an entropy.

The clustering algorithm's formulation presented here follows from two"claims."

Claim A: Given a set of data points $\{\vec{x}_i\}$ the optimal number of clusters is 1) the mostprobable number of clusters in 2) the maximum entropy membership distributionfor which 3) the average deviation from the average vanishes, i.e., $\langle \vec{x}_i - \langle \vec{x} \rangle \rangle = 0$ (averages$\langle\,\rangle$ are calculated over membershipdistributions).

The meaning of this claim is easily understood by noting that: point 1) refersto the number of clusters that is associated with the most probable resolution parameter, assuming the latter to be chosen randomly (i.e., in the absence of any information on the metric, there is no subjective bias introduced on the scale of the metric); point 2) refers to the fact that the probability distribution for a point to belong to a cluster must be a maximum entropy distribution (parameterized by the scale of the metric) to satisfy the criterion of lack of information (no bias); point 3) refers to the fact that the clusters should be as compact/uniform as possible.

To find the optimal position of the clustering centroids (after their number has been established) Claim A is supplemented with Claim B that is:

Claim B: The optimal positions of the cluster centroids correspond to theresolution which yields the minimum clustered entropy for all the partitions with the optimal number of clusters.

The meaning of this claim is that the clusters should be as separated aspossible and thus their centroids should be those that yield the minimum clustered entropy.

The only subjective element in the solution is in the form of the requirement of compactness, i.e., $\langle \vec{x}_i - \langle \vec{x} \rangle \rangle = 0$, which, more generally, could be written as $\langle d(\vec{x}_i, \langle \vec{x} \rangle) \rangle$ where $d(a, b)$ is a distance function not necessarily of the form $d(a, b) = a - b$.

Besides the low-bias, some other advantages of this algorithm are summarizedas follows. The solution gives us both hierarchy type clustering results and partitioning type clustering results; also the solution gives both hard and fuzzy partitions and a global optimal partition. The compactness/uniform and separation measurements are integrated in the clustering algorithm, but not used as the criterion to evaluate a partition's quality. The cost function is the clustered entropy which measures the inter cluster influences; thus, the cluster validity measurement results from our clustering procedures itself. Finally, the clustering-tendency test is implied in the algorithm. In constructing the validity diagram, we keep the cluster structure with one cluster to compete with other cluster structures. If it is the most valid, we can conclude that the data set is not well separated and consider it as being uniform.

References

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

[2] G. H. Ball and D. J. Hall, "Isodata, a novel method of data analysis and pattern classification," Stanford Res. Inst. Tech. Rep. (NTIS AD 699616), Stanford, CA, 1965.

[3] J. C. Bezdek, "Fuzzy mathematics in pattern classification," Ph.D. dissertation, Cornell Univ., Ithaca, NY, 1973.

[4] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–680, 1983.

[5] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and bayesian restoration of images," *IEEE Trans. Pattern. Anal. Machine Intell.*, vol. PAMI-6, pp. 721–741, 1984.

[6] K. Rose, E. Gurewitz, and G. C. Fox, "A deterministic annealing approach to clustering," *Pattern Recognit. Lett.*, vol. 11, pp. 589–594, Sept. 1990 (North-Holland).

[7] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley-Interscience, 1974.

[8] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 721–728, 1989.

[9] X.-L. Xie and G. Beni, "Validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, no. 8, pp. 841–847, Aug. 1991.

[10] B. S. Everitt, *Cluster Analysis.* New York: John Wiley, 1974.

[11] R. Dubes and A. K. Jain, "Clustering techniques: The user's dilemma," *Pattern Recognit.*, vol. 8, pp. 247–260, 1976.

[12] G. Matthews and J. Hearne, "Clustering without a metric," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, no. 2, pp. 175–184, Feb. 1991.

[13] E. T. Jaynes, "Information theory and statistical mechanics," in *Papers on Probability, Statistics, and Statistical Physics,* R. D. Rosenkrantz, Ed. Dordrecht, The Netherlands, Kluwer Academic Publishers, 1989 (Reprint of a 1957 paper).

[14] R. A. Fisher, "Multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.

[15] M. James, *Classification Algorithm.* New York: John Wiley, 1985, ch. 7.

[16] G. Beni and X. Liu, "An efficiency computation algorithm for the least biased fuzzy clustering method," in preparation.