A Maximum-Entropy Approach to Fuzzy Clustering

Rui-Ping Li and Masao Mukaidono
Department of Computer Science, Meiji University
Kawasaki-shi 214, Japan

Abstract — In this paper, we propose a new approach to fuzzy clustering by means of a maximum-entropy inference (MEI) method. The resulting formulas have the more beautiful form and the clearer physical meaning than those obtained by means of the fuzzy c-means (FCM) method. In order to solve the cluster validity problem, we introduce a structure strength function as clustering criterion, which is valid for any membership assignments, thereby being capable of determining the plausible number of clusters according to our subjective requisition. With the proposed stucture strength function, we also discuss global minimum problem in terms of simulated annealing. Finally, we simulate a numerical example to demonstrate the discussed approach, and compare our results with those obtained by the traditional approaches.

1. INTRODUCTION

Clustering methods are not only major tools to uncover the underlying structures of a given data set, but also promissing tools to uncover the local input-output relations of a complex system [1]. This naturally interests us and leads us to reinvestigate these methods. As well known, the clustering problem is an optimazation problem that a group of objects is split up into a plausible number of subgroups having some features on the basis of a measure functionan often subjectively chosen, such that the distance between objects within a subgroup is smaller than the distance between objects belonging to different subgroups. This is a very difficult problem because the well-defined knowledge is too little. Among the existing clustering methods, FCM method proposed by Bezdek [2] is one of the most active and often-used data analysis methods in recent years. This may be because of the widespread resurgence of interest in the theory and applications of fuzzy set and the generalized form of FCM method.

$$J_m = \sum_{k=1}^{N} \sum_{i=1}^{c} u_{ik}{}^m \|x_k - v_i\|^2 \quad 1 \le m < \infty. \qquad (1)$$

Now, let us reconsider this method. In FCM method the loss (objective) function is defined as shown eqn (1). Where, u_{ik} denotes the grade of membership of the k-th pattern in the i-th fuzzy cluster, v_i is interpreted as cluster center or prototype of the i-th cluster defined by u_{ik} (k=1 to N), and weighting exponent m controls the extent of membership sharing between fuzzy clusters. As m=1, FCM converges in theory to the traditinal k-means solution [3]. To minimize eqn (1) subject to normalization constraint, used Lagrangian multipliers method, for m>1, local minimum solutions of eqn (1) was demonstrated if and only if

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|} \right)^{2/(m-1)}} \qquad \forall\ i, k;$$

$$v_i = \sum_{k=1}^{N} u_{ik}{}^m x_k \Big/ \sum_{k=1}^{N} u_{ik}{}^m \qquad \forall\ i.$$

It is needless to say that, FCM is wonderful method because Bezdek introduced a "strange number" m, but it has still the following three open questions:

0-7803-2461-7/95/$4.00

1) What is the physical meaning of the paramter m ? and how does make an optimal choice of the paramter m ?
2) It discussed only local minimum solutions but do not mention how to get global minimum solutions.
3) There still exists no criterion which is valid for any membership assignments (contained *hard* clustering case) while has clear physical meaning.

On the other hand, from mathematical form of eqn (1), one may have felt that the exponent m coming on it is such unnatural and unnecessary.

Our work is inspired by the early work of Jaynes [4] in which he stated information theory and statistical mechanics in a sweeping form. In the following section of this paper, we first introduce a local loss function L, then derivate its solutions by means of maximum-entropy principle. The third section defines a *struture strength function* as our clustering criterion, and summarizes a Data Structure Recognition (DSR) algorithm. In the fourth section, we give a example to illustrate our method. Finally, the five section contains our conslusions.

2. MAXIMUM-ENTROPY INFERENCE IN FUZZY CLUSTERING

As haved been mentioned in preceding section, information theory (strictly speaking, Shannon's concept of "amount of informa-tion") provides an unbiased inference method for ill-defined problems on the basis of the given information, called a Maximum-Entropy Inference (MEI). The structure of the MEI problem is one of finding a probability assignment (or, membership function $\{u_{ik}\}$) which avoids bias, while agreeing with whatever information is given, each u_{ik} lies in [0,1] and that maximizes the entropy. Formally, this problem is written as:

$$maximize\left\{-\sum_{k=1}^{N}\sum_{i=1}^{c} u_{ik}\, log(u_{ik})\right\}, \qquad (2)$$

subject to:

c_1

c_2

.

c_m.

Where $c_1, c_2, \ldots, c_m$ are the m constraints or the given information.

Below, we will see that fuzzy clustering problem is just one of the MEI. The first, we define a loss function (the within-group sum-of-squared-error (WGSS)) as follows:

$$L = \sum_{k=1}^{N}\sum_{i=1}^{c} u_{ik} \cdot d_{ik}^2. \qquad (3)$$

Where, $d_{ik} = \| x_k - v_i \|$ and $\| \cdot \|$ is inner product, N is the number of data pairs, c is the number of clusters, x_k denotes the vector which represents the k-th data pairs, v_i is the vector which defines the centriod of the i-th cluster (or, prototype) and u_{ik} denotes the grade of membership of the k-th data pairs in the i-th cluster, and satisfys the following conditions:

$$0 \le u_{ik} \le 1 \qquad \forall\ i, k; \qquad (4a)$$

$$0 < \sum_{k=1}^{N} u_{ik} < N \qquad \forall\ i; \qquad (4b)$$

$$\sum_{i=1}^{c} u_{ik} = 1 \qquad \forall\ k. \qquad (4c)$$

Therefore, we see that with maximum-entropy inference, fuzzy clustering problem becomes one of finding a set of prototypes which minimize eqn (3) and a membership aasignment which satisfy constraints (4c). As it will be seen below, solutions which satisfy constraint (4c) also satisfy constraints (4a) and (4b). To maximize eqn (2) subject to the constraints (3) and (4c), we use Lagrangian multipliers method, and obtain the following solutions

$$u_{ik} = e^{-\frac{d_{ik}^2}{2\sigma^2}} / \sum_{j=1}^{c} e^{-\frac{d_{jk}^2}{2\sigma^2}} \quad \forall\, i,\, k. \qquad (5)$$

Where the parameter σ is Lagrangian multiplier determined eqn (3), called the *admissible error radius* in this paper. It is well-known that, $2\sigma^2$ is called the "temperature" in statistical physics.

Obviously, the resulting eqn (5) are solutions which satisfy the eqn (3) only when the centroid vectors $\{v_i\}$ were assumed to be constant vectors. As $\{u_{ik}\}$ fixed, we use the method of proof similar to [2], and obtain the following solutions which minimize eqn (3).

$$v_i = \sum_{k=1}^{N} u_{ik}\, x_k / \sum_{k=1}^{N} u_{ik} \qquad \forall\, i\,. \qquad (6)$$

This means that v_i is just the centroid of the i-th cluster.

Below we summarize the results of this section as a proposition.

Proposition 1. Assume that the N unlabeled data are given and c is fixed ($2<c<<N$).

1) If the membership assignments $\{u_{ik}\}$ are fixed, then the centroid vectors $\{v_i\}$ are solutions which minimize locally eqn (3) if and only if

$$v_i = \sum_{k=1}^{N} u_{ik}\, x_k / \sum_{k=1}^{N} u_{ik} \qquad \forall\, i\,. \qquad (7)$$

2) If the centroid vectors $\{v_i\}$ are fixed, then the membership assignments $\{u_{ik}\}$ which satisfy eqn (3) are just Gaussian distributions as follows

$$u_{ik} = e^{-\frac{d_{ik}^2}{2\sigma^2}} / \sum_{j=1}^{c} e^{-\frac{d_{jk}^2}{2\sigma^2}} \quad \forall\, i,\, k. \qquad (8)$$

Proof: As stated in this section.

3. STRUCTURE STRENGTH AND A DATA STRUCTURE RECOGNITION ALGORITHM

So far we only discussed such a case in which the number of clusters c is fixed, but do not mention how to compare two divisions with the different number of clusters. This problem is called the *cluster validity problem* and studied widely in [2],[5] and [6]. Among proposed criterions, the *partition coefficient*, the *partition entropy* [2] and the *new method* [5] are methods which are often used. But all of these criterions depend strictly on the parameter m, that is, these will be *invalid* when the amount of fuzziness (or, m) is either large or small (usually, $1.5<m<3.0$).

In this paper, we introduce a new concept, which is called *structure strength*. The existence of data structure means that the knowledge of a part allows us to guess easily the rest of the whole. Thus, the process of structure recognition amounts to a process of the knowledge extraction. Therefore the structure strength of a system will be very naturally expressed by the following formula

S = structure strength
= (the effectiveness of the classification)
+ (the accuracy of the classification)

For instance, if the number of data in input is N, and the number of clusters and the average total loss are c and $L(c)$ respectively, then the strength will be

$$S(c) = \alpha E + (1-\alpha)A$$
$$= \alpha \log(N/c) + (1-\alpha)\log(L(1)/L(c)). \qquad (9)$$

Where $L(1)$ is the variance of the entire data, i.e., being value of loss function as $c = 1$, and $\alpha \in [0,1]$ presents the "weight" of the classification effectiveness. In unbiased estimation, we make $\alpha = 0.5$. The first term (or, called the *information compression rate*) decreases with the number of clusters c, but the second term increases, because $L(c)$ *monotonously* decrease to zero with c. On the other hand, clustering, as

an approach of structure recognition, its objective is to find the strongest structure. For cluster validity, we thus consider maximization of $S(c)$ as the clustering criterion.

$$Max\{S(c),\ c=1\ to\ c\}. \tag{10}$$

Because the *structure strength* is defined by a logarithmic function, the strength is *additive*. In this sense, if the transition from $M = M_0$ (M_0=N denotes the number of data) to $c = M_k$ is done through successive steps: $M_0 \rightarrow M_1 \rightarrow M_2 \rightarrow \cdots$, $\rightarrow M_k$, we have

$$S = \sum_{i=1}^{k} S_i$$

$$S_i = \alpha \log \frac{M_{i-1}}{M_i} + (1-\alpha) \log \frac{L(M_{i-1})}{L(M_i)} \tag{11}$$

but $L(N) \equiv L(1)$, because we have the following constraint condition : $2 \leqq c \ll N$.

Therefore, our method is called the Data Structure Recognition (DSR). DSR algorithm is composed of two steps:

STEP 1 calculate the loss function $L(c)$ for fixed c. It is therefore necessary to find the global minimization of eqn (3).

STEP 2 determine the plausible number of clusters by varying c.

In order to perform the minimization required in STEP 1, DSR algorithm is based on *Proposition 1*, i.e., to use eqn (7) and eqn (8) repeatedly until a termination criterion. On the other hand, Geman and Geman [8] have proofed that, in theory, the global minimization of eqn (3) can be achieved if the computatinal temperature is inversely propor-tional to a logarithmic function of time. Therefore, if necessary, for fixed c, an optimal choice of σ can be made in terms of simulated annealing. But, such cooling schedules are very slow and unrealistic in many applications. However, it is very fortunate that a complete cooling schedule dose not need for clustering problem, since $L(c)$ is strictly depented on the number of clusters c when the admisible error radius σ is small. As σ gets smaller, the membership assignments become less fuzzy. As $\sigma = 0$, it becomes a *hard* clustering problem. As $\sigma \rightarrow \infty$, each pattern is equally assignmented to all clusters, and thus we get only a *single* cluster. Therfore, σ not only determines value of loss function L, but also the number of clusters c. From this meaning, here σ has similar properties to m in [2]. But σ has clearer physical meaning than m, and an optimal choice in terms of simulated annealing, whereas m has not.

Since L is depend on c, STEP 2 is necessary, i.e., finding the maximization of $S(c)$ by varying c to determine the plausible number of clusters. We summarize the DSR algorithm as follows:

DSR algorithm:

1) Compute $L(1)$. Fix $\sigma > 0$, $\varepsilon > 0$, α, C, T and $S(1) = 0.0$.
2) For $c=2,3,\ldots,C$. Initialize $u_{ik} \in [0,1]$ at random for each i, k. For $t=1,2,\ldots,T$.
3) Calculate $\{v_i(t)\}$ using (7) and $\{u_{ik}\}$.
4) Update $\{u_{ik}(t)\}$ using (8) and $\{v_i\}$.
5) IF $max \parallel u_{ik}(t)-u_{ik}(t-1) \parallel > \varepsilon$ next t; ELSE calculate $S(c)$ using (9).
6) IF $S(c)<S(c-1)$ stop; ELSE next c.

4. ILLUSTRATIVE EXAMPLE

In this section, we give a numerical example to illustrate our method. It is an artificial data set of N=36 points. For the sake of intuition, we have chosen a two-dimensional data distribution as shown in Fig.1. We applied DSR algorithm to it with $\sigma = 0.7$ and $\varepsilon = 0.001$, $\alpha = 0.5$ and found the plausible number of clusters $c^* = 3$. For comparison, we also applied FCM algorithm to it with $m = 2.0$ and $\varepsilon = 0.001$, and get a identical result: $c^* = 3$. But, the valuel of loss function obtained by DSR method is smaller (about one third) than that by FCM method, and

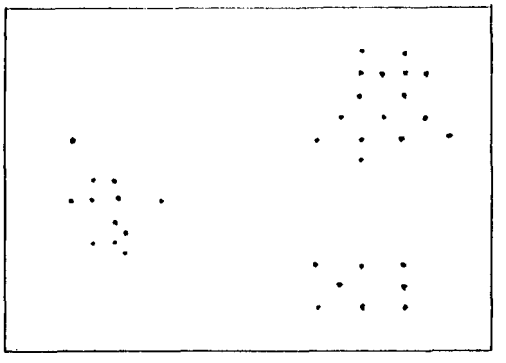

Fig. 1 An artifical two-dimensinal data set

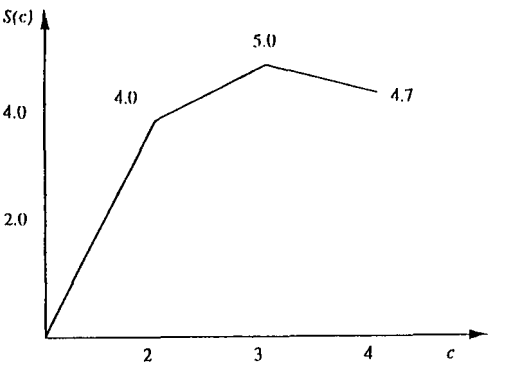

Fig. 2 validity indicator for $S(c)$

Table 1 Membership assignments at c=3. From FCM, centers of clusters are (0.23, 0.33), (0.80, 0.15) and (0.85, 0.58); value of loss function $J(3)$=0.973. From DSR, centers of clusters are (0.23, 0.33), (0.80, 0.15) and (0.84, 0.57); value of loss function $L(3)$=0.323.

Data	From FCM			From DSR		
h	(0.23, 0.33)	(0.80, 0.15)	(0.85, 0.58)	(0.23, 0.33)	(0.80, 0.15)	(0.84, 0.57)
1	0.971	0.015	0.013	1.000	0.000	0.000
2	0.880	0.056	0.062	1.000	0.000	0.000
3	0.966	0.019	0.013	1.000	0.000	0.000
4	0.993	0.003	0.003	1.000	0.000	0.000
5	0.974	0.013	0.012	1.000	0.000	0.000
6	0.965	0.020	0.014	1.000	0.000	0.000
7	0.993	0.003	0.002	1.000	0.000	0.000
8	0.996	0.001	0.001	1.000	0.000	0.000
9	0.974	0.013	0.012	1.000	0.000	0.000
10	0.929	0.043	0.027	1.000	0.000	0.000
11	0.972	0.016	0.011	1.000	0.000	0.000
12	0.906	0.051	0.041	1.000	0.000	0.000
13	0.047	0.901	0.050	0.000	1.000	0.000
14	0.050	0.879	0.070	0.000	1.000	0.000
15	0.010	0.973	0.016	0.000	1.000	0.000
16	0.007	0.979	0.012	0.000	1.000	0.000
17	0.006	0.979	0.014	0.000	0.999	0.000
18	0.021	0.931	0.046	0.000	1.000	0.000
19	0.017	0.938	0.043	0.000	1.000	0.000
20	0.021	0.915	0.067	0.000	0.999	0.000
21	0.088	0.166	0.745	0.000	0.000	0.999
22	0.031	0.062	0.905	0.000	0.000	1.000
23	0.045	0.174	0.780	0.000	0.000	0.999
24	0.023	0.068	0.908	0.000	0.000	0.999
25	0.007	0.014	0.978	0.000	0.000	1.000
26	0.016	0.028	0.954	0.000	0.000	1.000
27	0.033	0.051	0.914	0.000	0.000	1.000
28	0.002	0.005	0.992	0.000	0.000	1.000
29	0.009	0.019	0.971	0.000	0.000	1.000
30	0.017	0.063	0.919	0.000	0.000	0.999
31	0.005	0.013	0.981	0.000	0.000	1.000
32	0.012	0.027	0.959	0.000	0.000	1.000
33	0.026	0.050	0.922	0.000	0.000	1.000
34	0.026	0.101	0.872	0.000	0.000	0.999
35	0.016	0.040	0.939	0.000	0.000	1.000
36	0.017	0.063	0.919	0.000	0.000	0.999

membership assignments are less fuzzy (Table 1). Here, we consider eqn (9) as clustering criterion, behavior of $S(c)$ is shown in Fig. 2. It is worth emphasis that the existing clustering criterions, e.g., the *partition coeficient*, the *partition entropy* [2] and the *new method* [5] all are invalid for such less fuzzy membership assignments.

5. CONCLUSIONS

In this paper we have shown that a complete Data Structure Recognition (DSR) method can be derived by means of the maximum-entropy inference. The characteristic paramters of the method are the *admissible error radius* σ and the *structure strength function* S. As compared with FCM method, our method exhibits following advantages: 1) having clearer phy-sical meaning and well-defined mathematical features; 2) having an optimal choice for σ, and thus having high accuracy (or, less value of loss function); 3) having a clustering criterion which is *valid* for any membership assignments. We have given a example to illustrate our method. It has also been shown that our method not only have complete theoretical fundament but is very valuable in application.

REFERENCES

[1] R.P. Li and M. Mukaidono, "Fuzzy modelling and clustering neural network", 3rd International conference on Fuzzy Logic, Neural Nets and Soft Computing *IIZUKA'94*, JAPAN, pp.625-628, 1994.

[2] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, *New York,Plenum*, 1981.

[3] J. MacQueen, "Some methods for classification and analysis of multivariate observations", in *proc. 5th Berkeley symp. Mathematical statist. and Probability*, pp. 281-297, 1967.

[4] E.T. Jaynes, "Information theory and statistical mechanics", *Phys. Rev.* vol.106, pp. 620-630, 1957; vol.108, pp. 171-190, 1957.

[5] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for fuzzy c-means method", in *proc. 5th Fuzzy system symposium* (in Japanese), pp. 247-250, 1989.

[6] J.C Dun, "Well-separated cluster and optimal fuzzy partition", *J. Cybern.*, Vol. 4, pp. 95-104, 1974.

[7] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of image", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 721-741, 1984.