PROBABILISTIC SETS IN EVALUATION OF FUZZY CLUSTERING ALGORITHMS

Hajime Hirota

Department of Instrument and Control Engineering
College of Engineering
Kyushu University
Fukuoka 184, Kajino-cho 3-7-2
Japan

Witold Pedrycz

Department of Automatic Control and Computer Sci.
Silesian Technical University
Gliwice 44-100
Poland

Abstract The paper points out an application of probabilistic sets, especially their entropy measures, in evaluation of fuzzy clustering algorithms. We discuss several properties of the structure detected by the clustering algorithms.

Introduction

Unsupervised pattern recognition (clustering techniques) can be discussed as an important tool for discovering a structure of a data set under consideration [1][5][6]. The clustering algorithm generates partition of the entire data set $X=\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$, $\underline{x}_i \in \mathbb{R}^m$, into subsets such that more or less homogenous elements of X are assigned to the same subset. In general, the c-clusters obtained may be characterized in a unique way by a set of "c" disjoint subsets of X: $S_1, S_2, \ldots, S_c$, that union is X,

$$X = \bigcup_{i=1}^{c} S_i \,. \tag{1}$$

An equivalent way of defining the clusters is obtained using characteristic functions of S_i's

$$\chi_i : X \longrightarrow \{0,1\} \,. \tag{2}$$

Of course, due to the properties of subsets S_i specified above, one gets,

$$\bigwedge_{1 \le j \le n} \sum_{i=1}^{c} \chi_i(\underline{x}_j) = 1 \qquad \underline{x}_j \text{ belongs to exactly one cluster} \tag{3}$$

$$\bigwedge_{1 \le i \le c} \sum_{j=1}^{n} \chi_i(\underline{x}_j) > 0 \qquad \text{each cluster is a nonempty subset of X} \tag{4}$$

From a formal point of view, fuzzy sets allow us to relax the constraints shown before, viz. every $\underline{x}_j$ may belong to the cluster specified with a certain grade within [0,1] interval. This grade reflects a strength of belonginess of the element of X to the cluster. Thus we can speak the clusters are fuzzy indeed, and therefore it is possible to analyze the data in more extended fashion. An advantage of the use of fuzzy sets is twofold:

- firstly, we have at our disposal a deep insight into a structure of the data set (grades of belonginess of the element to the cluster) that allows the user to **compare the most significant** substructure in X and a residual one,
- secondly, fuzzy sets offer us a tool for selecting the most "plausible" number of clusters in the structure discussed (for exhaustive discussion the reader is referred e.g. to [2]-[4], [10], [11]).

Similarly as in the situation while one applies hard clustering techniques, we have to look for an effective mechanism of interpretation of the results of fuzzy clustering. For hard clustering one usually prefers cross-partition method (cf. [1]). In this case for

fuzzy clustering, the problem of evaluation of the results has not been completely solved. Remembering the fact the clustering forms an initial stage for data processing and creates a fedback loop between the clustering mechanis algorithms and a data-analyst, the problem of evaluation of the results obtained is a crucial one [5].

Nowadays the number of fuzzy clustering algorithms is rather large. Take for instance: Fuzzy c-means (an extended version of ISODATA)[3], Fuzzy c-varietes [3] or generally geometrical fuzzy clustering [11], clustering with dissimilarity measures (cf. [2] [9]). Each of them produces a partition of the data set X. Moreover some of them are preferrable with respect to the shape of the clusters in X. Nevertheless, in general, one is interested in evaluation of the results of the overall analysis.

In this paper we shall present the way how the results of clustering can be interpreted in terms of probabilistic sets [7] and ensembles of probabilistic sets. Due to extensive analysis of this generalization of fuzzy sets which exists in literature (see [7] [8]) we shall restrict ourselves to applicational aspects of probabilistic sets in clustering.

Evaluation of the results of the fuzzy clustering

Let us start with the results of the fuzzy clustering methods collected in a form of a partition matrix $U = [u_{ij}]$, $i=1,2,\ldots,c$, $j=1,2,\ldots,n$. According to the previous discussion U fulfills a collection of corditions,

$$(i) \quad \forall_{1\le i\le c} \ \forall_{1\le j\le n} \ u_{ij}\in[0,1] \tag{5}$$

$$(ii) \quad \forall_{1\le j\le n} \ \sum_{i=1}^{c} u_{ij}=1 \tag{6}$$

$$(iii) \quad \forall_{1\le i\le c} \ \sum_{j=1}^{n} u_{ij}>0 \tag{7}$$

It is obvious that every column of U represents grades of belonginess of the elements in X to the concrete class;say

$$\underline{u}_1=[u_{11}\ u_{12}\dots u_{1n}] \tag{8}$$

is the membership function of the fuzzy set $\underline{u}_1$.

When X is analyzed by "M" clustering methods,we obtain "M" various partition matrices $U^1,U^2,\dots,U^M$.First of all we are interested in establishing a correspondence between the clusters generated by various algorithms.This step may be performed in diverse fashions.A foundation of one of the method lies in comparison distances between rows of U^j, s.The assignment leads us to "c" tables containing the values of the membership functions of the clusters in X. They take the following form,

clustering method / element of X	method 1	method 2	....	method M
$\underline{x}_1$	u^1_{m1}	u^2_{m1}	....	u^M_{m1}
$\underline{x}_2$	u^1_{m2}	u^2_{m2}	....	u^M_{m2}
$\vdots$				
$\underline{x}_n$	u^1_{mn}	u^2_{mn}	....	u^M_{mn}

Table for the m-th cluster (m=1,2,...,c)

Thus we have "c" probabilistic sets $U_1,U_2,\dots,U_c$ that are represented in the tables shown above.

Let us discuss the representation of these results in terms

of entropy of probabilistic sets (cf.[8]). The complexity of the object, say $\underline{x}_j$, with respect to the category U_i is given by entropy $H(\underline{x}_j, U_i)$. The mutual entropy between U_i, U_l for $\underline{x}_j$ is given by $H(\underline{x}_j, U_i, U_l)$. A level of interactivity between U_i and U_l for $\underline{x}_j$ is indicated by $I(\underline{x}_j, U_i, U_l)$ equal to sum of entropies $H(\underline{x}_j, U_i)$ and $H(\underline{x}_j, U_l)$ diminished by a mutual entropy $H(\underline{x}_j, U_i, U_l)$:

$$I(\underline{x}_j, U_i, U_l) = H(\underline{x}_j, U_i) + H(\underline{x}_j, U_l) - H(\underline{x}_j, U_i, U_l) \qquad (9)$$

The abovementioned indices are of a local character due to the fact a concrete object $\underline{x}_j$ is taken into account. The global indices representing the properties of the overall structure may be of a great help,

$$\bar{H}(X, U_i) = 1/n \sum_{j=1}^{n} H(\underline{x}_j, U_i) \qquad (10)$$

$$\bar{H}(X, U_i, U_l) = 1/n \sum_{j=1}^{n} H(\underline{x}_j, U_i, U_l) \qquad (11)$$

$$\bar{I}(X, U_i, U_l) = 1/n \sum I(\underline{x}_j, U_i, U_l) \qquad (12)$$

They convey an information about complexity of the data structure discovered by the clustering methods and interactivity between the clusters. If interactivity is high enough, we may suspect that the data are not well separated or the clustering methods are not appropriate and cannot detect the structure in X.

Numerical illustration

In order to visualize the performance of the entropy of the probabilistic sets for unsupervised classification let us consider data set shown in Fig.1

consists of 18 points in R^2

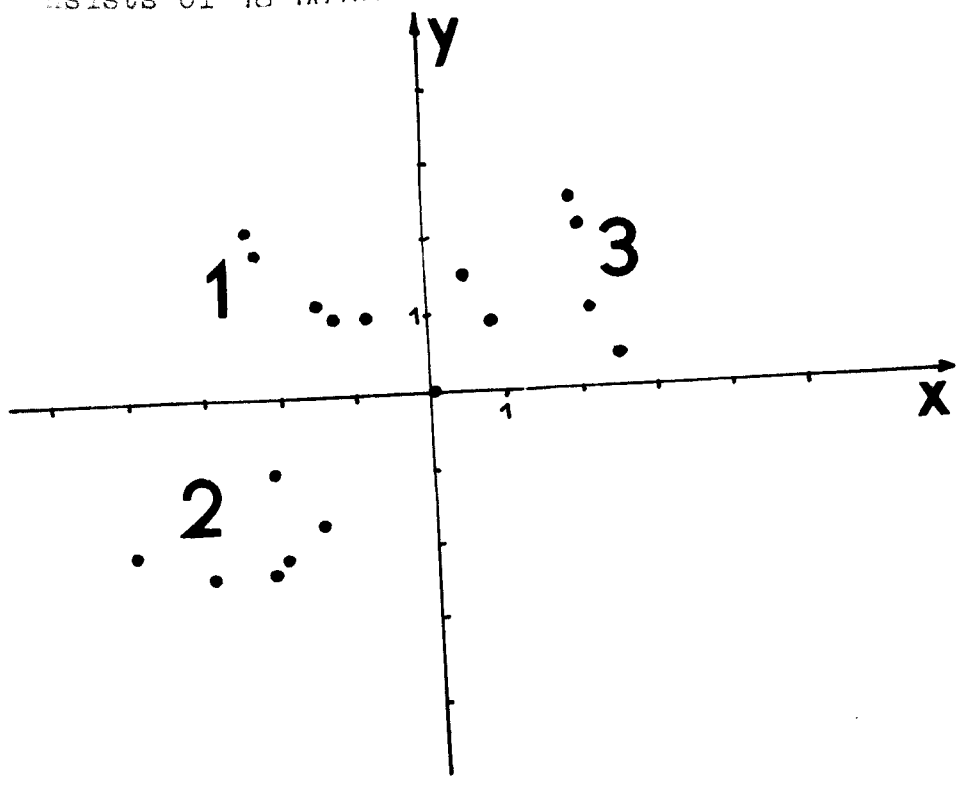

Fig.1.Two dimensional data set with three clusters detected

Three clustering algorithms were used:

(1) FUZZY ISODATA with p=1.5

(2) FUZZY ISODATA with p=2.5

(3) Roubens' method[9]with the dissimilarity measure between the elements treated as Euclidean distance.

Starting with c=3 we have the following values of the entropies of the clusters:

$\bar{H}(X,U_1)=1.595$, $\bar{H}(X,U_2)=1.68$, $\bar{H}(X,U_3)=1.52$

and

$\bar{I}(X,U_1,U_2)=0.87$, $\bar{I}(X,U_1,U_3)=0.83$, $\bar{I}(X,U_2,U_3)=0.85$.

Note that the complexity of all the clusters is the same;similarly

interaction between the pairs of the clusters does not vary in X.

Afterwards, the data set X is forced to be splitted into two

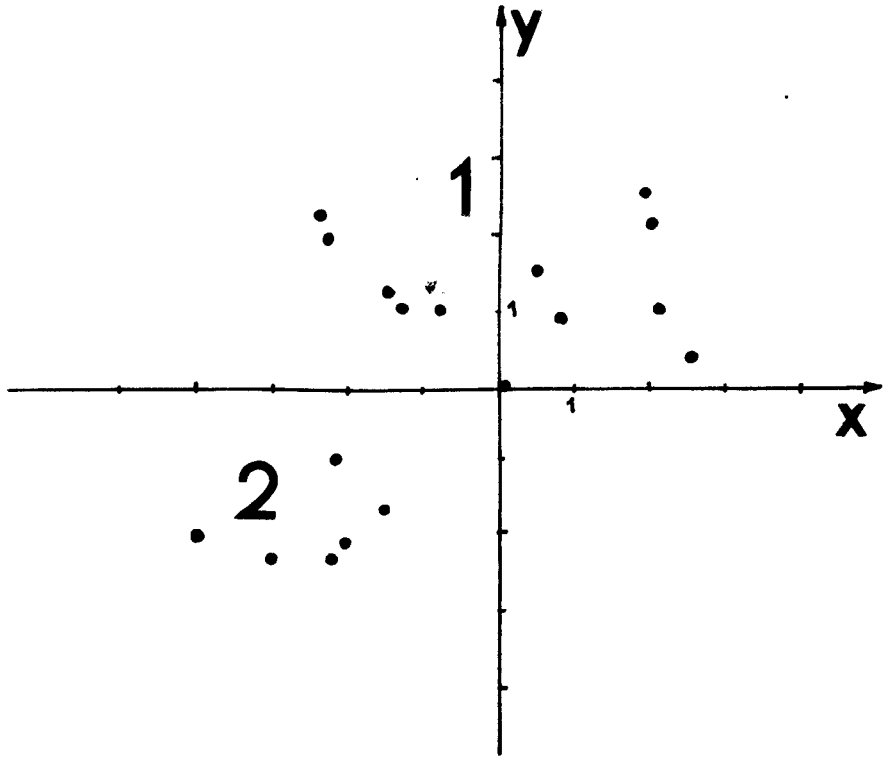

Fig.2. Two dimensional data set with two clusters detected

Here one gets,

$\bar{I}(X,U_1)=2.07$, $\bar{I}(X,U_2)=2.03$ $\bar{I}(X,U_1,U_2)=1.36$.

Moreover, a hypothesis appears that the measure of interactivity permits us to use it for detecting the most"plausible" number of clusters. For this purpose consider an averaged interactivity between the clusters,

$$\tilde{I}= 1/L \sum_{i,j:i<j} \bar{I}(X,U_i,U_j) \qquad (13)$$

The sum is taken over all the pairs of i,j=1,2,...,c such that i<j. The number of pairs is equal to L, $L= c(c-1)/2$. The plot $\tilde{I}$ versus c is depicted in Fig.3.

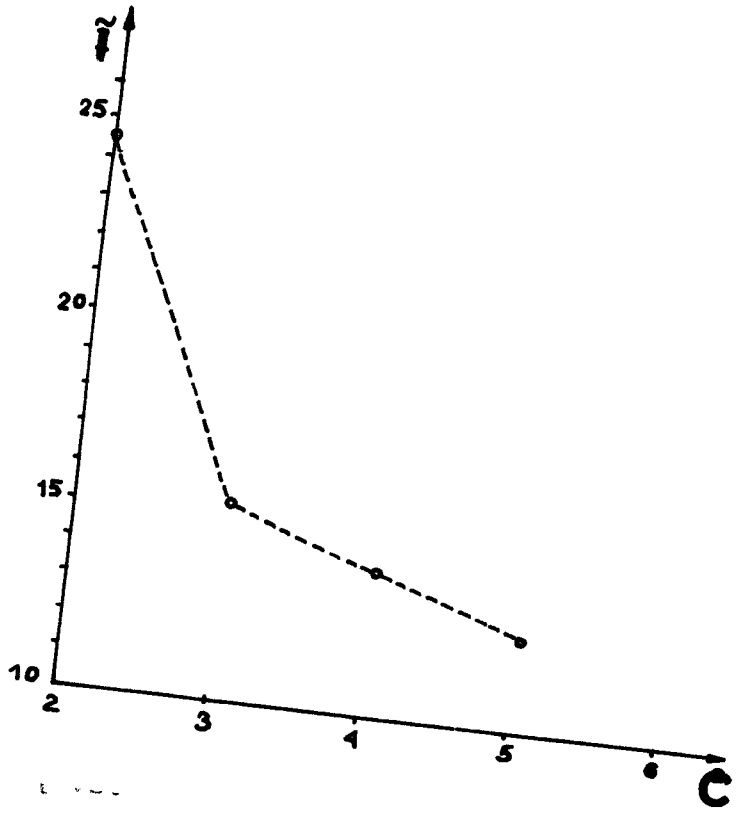

Fig. [illegible]

By inspection, $\tilde{I}$ is a decreasing function of c. A significant jump in values of $\tilde{I}$ for a certain "c" suggests us to treat it as the most "plausible" number of groups. Here, the jump occurs while "c" changes from 2 to 3, that suggests c=3 clusters detected in X.

References

[1] M.R. Anderberg, Cluster Analysis for Applications. Academic Press, New York, 1973.

[2] E. Backer, Cluster Analysis by Optimal Decomposition of Induced Fuzzy Sets, Delft Univ. Press, Delft, 1978.

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

[7] J.C.Bezdek,I.P.Windham,R.Ehrlich,Statistical parameters of cluster validity functionals,Int.J.Computer Information Sciences, 9,1980,323-336.

[8] J.Dubes,A.Jain,Validity studies in clustering methodologies, Pattern Recognition,11,1979,235-253.

[9] R.O.Duda,P.E.Hart,Pattern Classification and Scene Analysis, Wiley,New York,1973.

[10] M.Hirota,Concepts of probabilistic sets,Fuzzy Sets and Systems, 7,1981,31-46.

[11] M.Hirota,Ambiguity based on the concept of subjective entropy. In:Fuzzy Information and Decision Processes(M.M.Gupta,E.Sanchez, Eds.) North-Holland,Amsterdam,1982,pp.29-40.

[12] H.Roubens,Pattern classification problems and fuzzy sets,Fuzzy Sets and Systems,1,1978,239-253.

[13] M.P.Windham,Cluster validity for the fuzzy c-Means clustering algorithm,IEEE Trans.PAMI,4,1982,357-363.

[14] M.P.Windham,Geometrical fuzzy clustering algorithms,Fuzzy Sets and Systems,10,1983,271-279.