Fuzzy Clustering Level Analysis Applying AIC

Shuya Kanagawa[1] (Musashi Institute of Technology, Tokyo, Japan)
Ei Tsuda (Kokugakuinn SR.HS. , Tokyo, Japan)
Kimiaki Shinnkai (Waseda University, Tokyo, Japan)
Hajime Yamashita (Waseda University, Tokyo, Japan)

1. Introduction

As for analysis of inexact information such as human behavior, mental process, social structure and so on, we consider fuzzy graph of some relation in these problems. Fuzzy graph is constructed with clusters of some levels. Tsuda and Yamashita (1994) proposed a rational method to decide the optimal level of fuzzy clustering with partition tree. In this paper we improve his method to use AIC (Akaike's information criterion) which is a likelihood estimator via Kulback-Leibler information.

2. Tsuda-Yamashita method for fuzzy clustering level analysis

Clustering level analysis of fuzzy graph is useful to analize inexact information such as human behavior, mental process, etc, see e.g. Romsburg (1984). It is difficult to decide the optimal level of fuzzy clustering as to a partition tree.

Fig.1 is an example of fuzzy clustering with partition tree. As for Fig.1 the set of clustering levels is {0.00, 0.13, 0.27, 0.49, 0.58, 0.59, 0.74, 0.79, 0.91, 1.00} and the optimal level of the partition tree in the set. For example the optimal level 0.58 means that the set of ten points {1,2,…,10} is constructed from 5 clusters {6}, {3}, {2,7,1,4,9}, {5,8}, {10}. In 1994 Tsuda and Yamashita suggested a new method to obtain the optimal level to find a unique equilibrium point for the cluster number and the cluster size. It means to know a kind of stable classification of clusters for {1,2,…,10}. There are many applications obtained by their method which are fit well for actual examples. In this note we improve their method to apply the AIC method.

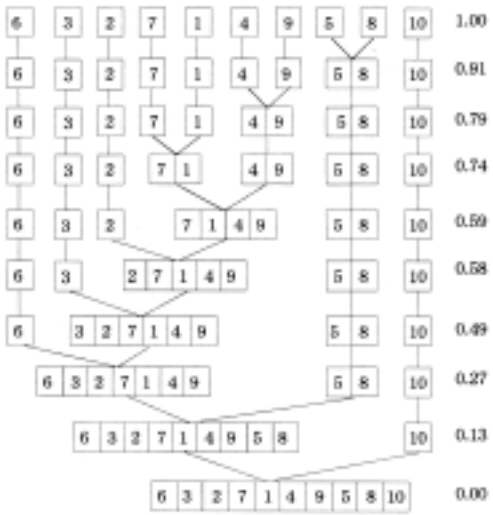

Fig.1. Fuzzy clustering with partition tree

3. AIC (Akaike's information criterion)

AIC is an identification method using Kullubback-Leibler information numbers as well as the consistency and the asymptotic normality of maximum likelihood estimators. Let X_1, X_2, …, X_m be independent and identically distributed random variables with probability density function $f(x,\theta)$, where θ is a parameter of the distribution. Further θ_k is a maximum likelihood estimator defined via Kullubback-Leibler information numbers and k is the dimension of the space in which θ exists. AIC for $f(x,\theta)$ is defined by

$$AIC(k) = -2\sum_{i=1}^{m} \log f\left(x,\theta_k\right) + 2k.$$

Comparing several models to fit the sampling data X_1, X_2, …, X_m, the smallest AIC of each model means the optimal one.

4. Fuzzy clustering level analysis with AIC method

As a typical example we treat Fig. Focusing to compare two clustering levels 0.74 and 0.59 we explain how to

[1] Research supported in part by Grant-in-Aid Scientific Research (No. 16540124), Ministry of Education, Science and Culture.
E-mail:kanagawa@ma.ns.musashi-tech.ac.jp

analize Fig.1 applying AIC method. The difference between two clustering levels is, in concrete terms, {7,1,9,4} is one cluster or is a combination of two clusters {7,1} and {9,4}.

Step 1. Configuration of {1,2,…,10}

From the fuzzy clustering levels find a configuration on the interval of all points. Since {5, 8} is a cluster at the level 0.91, the difference between {5} and {8} is 1-0.91=0.09. {4,6} is a cluster at the level 0.79 and the difference is 1-0.79=0.21, and so on (Fig.2). Then {1,2,…,10} has a configuration on the interval [0,8.54].

Step 2. Partitions of interval [0,8.54]

Divide the interval [0,8.54] into 3 or 6 as follows (Fig.3, Fig.5). Fig.3 implies that {7,1,9,4} is one cluster and Fig.5 suggests {7,1} and {9,4} are separated.

Step 3. Histograms for Fig.3 and Fig.5 (Fig.4, Fig.6).

Consider two histograms Fig.4 and Fig.6 are models of distribution which fits {1,2,…,10}. Then we can find which histogram is better to calculate their AIC.

Step 4. Calculate AIC for each histogram

$$AIC(Fig.4) = -2\left\{\log\frac{1}{20} + 6\log\frac{6}{20} + 3\log\frac{3}{10}\right\} + 2(3-1) = 35.8219$$

$$AIC(Fig.6) = -2\left\{\log\frac{1}{10} + 2\log\frac{2}{10} + 2\log\frac{2}{10} + 2\log\frac{2}{10} + 2\log\frac{2}{10} + \log\frac{1}{10}\right\} + 2(5-1) = 37.4162$$

Since AIC(Fig.4}< AIC(Fig.6} the distribution of the cofiguration of (1,2,…,10) on the interval [0,8.54] fits Fig.4 better than Fig.6. Therefore we conclude that the clustering level 0.59 is better than 0.74.

References

[1] H. Akaike, A new look at the statistical model identifications, IFFF Trans. Automatic Control, AC-19, 716--723.

[2] C.Romsburg : Cluster Analysis for Researchers Lifetime Learning Publication,1984.

[3] E.Tsuda and H. Yamashita, Sociometry analysis using fazzy graphs,J. Japan SOFT, vol. 6 , No.3, 1994 (in Japanese).

[4] E.Tsuda, A. Yanai and H. Yamashita, Decision Analysis of Optimal Fuzzy Cluster Level, Asian Fuzzy Systems Symposium Ⅳ 1T2E-4, 2000.

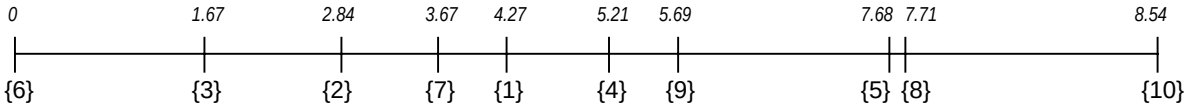

Fig. 2

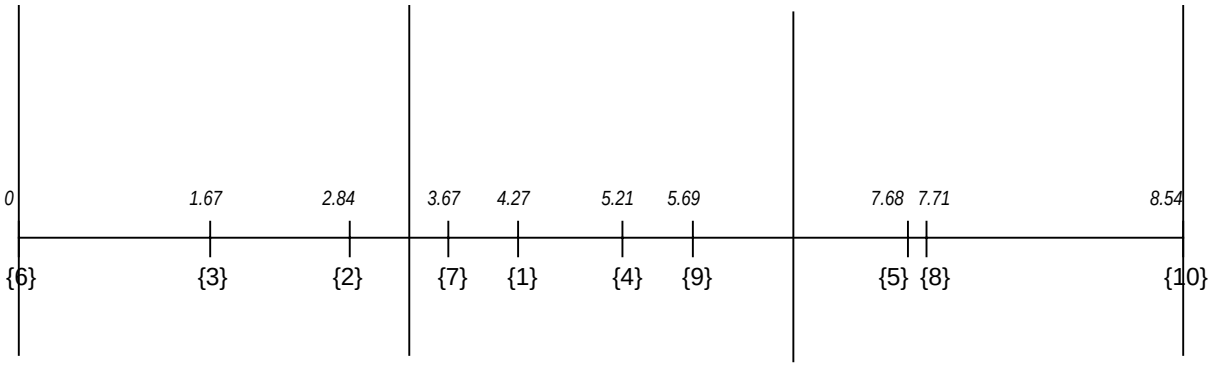

Fig. 3

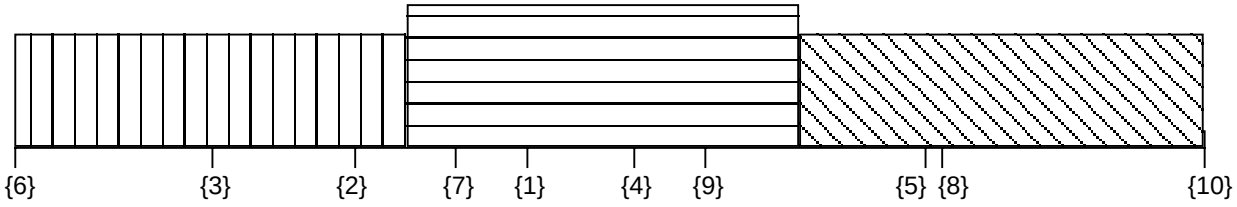

Fig. 4

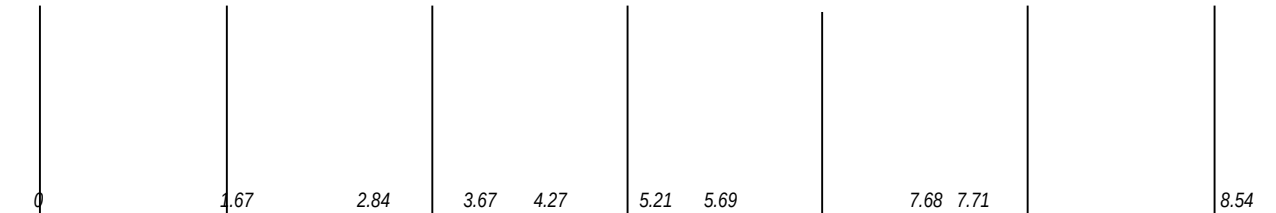

Fig. 5

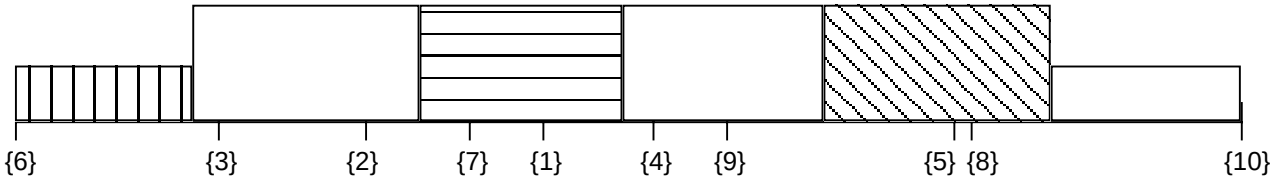

Fig. 6