Cluster Validity and Stability of Clustering Algorithms*

Jian Yu, Houkuan Huang, and Shengfeng Tian

Dept. of Computer Science, Beijing Jiaotong University
Beijing 100044, P.R.China
jianyu@center.njtu.edu.cn

Abstract. For many clustering algorithms, it is very important to determine an appropriate number of clusters, which is called cluster validity problem. In this paper, we offer a new approach to tackle this issue. The main point is that the better outputs of clustering algorithm, the more stable. Therefore, we establish the relation between cluster validity and stability of clustering algorithms, and propose that the conditional number of Hessian matrix of the objective function with respect to outputs of the clustering algorithm can be used as cluster validity cluster index. Based on such idea, we study the traditional fuzzy c-means algorithms. Comparison experiments suggest that such a novel cluster validity index is valid for evaluating the performance of the fuzzy c-means algorithms.

1 Introduction

Cluster analysis plays an important role in pattern recognition fields. However, the outputs of clustering algorithms are sensitive to parameters of the clustering algorithm. Sometimes, the same algorithm can lead to totally different outputs with respect to different parameters. A good clustering algorithm could produce undesirable results if parameters are chosen improperly. In the literature, many researches have been done on how to choose the optimal parameters in the clustering algorithms, particularly on how to choose the optimal number of clusters, for example, [1-4]. In general, selection of appropriate number of clusters and evaluation of outputs of clustering algorithms are called cluster validity problem.

In the literature, how to choose the optimal number of clusters depends on specific clustering algorithm. Many results on this issue are relevant to *c*-means or fuzzy c-means, for example, [1-4]. As for cluster validity for the FCM, one common approach is to design a cluster validity index to evaluate the performance of clustering algorithms. Frequently, there are two ways to design cluster validity index. One is based on the concept of fuzzy partition, the main assumption is that the performance of the FCM is better when its outputs are closer to crisp partition, for example, partition coefficient [2], partition entropy [5], uniform data functional [6], proportion

* This work was partially supported by the Key Scientific Research Project of MoE, China under Grant No.02031 and the National Natural Science Foundation of China under Grant No.60303014.

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 957–965, 2004.

exponent [7], nonfuzziness index [8], *etc*. However, as noted in [9], they lack of direct connection to the geometrical property of data set. Taking into account geometrical property results in another way to design cluster validity index, for instance, Xie-Beni index [9], Gunderson's separation coefficient [10], *etc*. In fact, the above cluster validity indices have the similar drawback, i.e., all of them do not pay enough attention to the property of the FCM itself.

From a point of algorithmic view, it is necessary to study the properties of clustering algorithms in order to determine number of clusters and evaluate clustering results.

Speaking roughly, given that data set truly follows the assumption of clustering algorithm, the probable outputs of a clustering algorithm should be the optimal clustering results. Obviously, it is a reasonable assumption. Otherwise, it has little chance to obtain the optimal clustering results no matter what cluster validity index is used. Therefore, we need to measure the probability of occurrence with respect to different outputs produced by clustering algorithm. It is easy to conjecture that clustering results with high stability are outputted with large probability. Therefore, we need to obtain and study the stability criterion of clustering algorithm. In the following, we study cluster validity for the FCM according to the above idea.

The reminder of this paper is organized as follows: In Section 2, the FCM and relevant cluster validity indices are related. In Section 3, a new cluster validity index, stable index, is defined and analyzed. In Section 4, numerical experiments are carried out to make a comparison between Xie-Beni index and our cluster index, and experimental results are analyzed. In the final, we draw conclusion and make a discussion.

2 The FCM Algorithm and Related Cluster Validity Indices

Let $X=\{x_1,x_2,\ldots,x_n\}$ be a s-dimensional data set, $u=\{u_{ik}\}$ is partition matrix, $v=\{v_1,v_2,\ldots,v_c\}$ is clustering prototype, the objective function is defined as $J_m(u,v,X)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m \| x_k - v_i \|^2$. The aim of the FCM algorithm is to obtain the partition matrix $u=\{u_{ik}\}_{c\times n}$ and clustering prototype $v=\{v_1,v_2,\ldots,v_c\}$corresponding to the minimum of the objective function J_m, where $\forall k \in \{k \mid 1 \le k \le n\}$, $\forall i \in \{i \mid 1 \le i \le c\}$, the membership u_{ik} represents the degree that x_k belongs to the clustering center v_i, and $u=[u_{ik}]_{c\times n} \in M_{fcn}=\{ u=[u_{ik}]_{c\times n} \mid \sum_{i=1}^{c} u_{ik}=1, u_{ik} \ge 0, 0<\sum_{i=1}^{c} u_{ik} < n\}$.

By Lagrange multiplier's approach, we obtain the necessary conditions for the minimum of $J_m(u,v,X)$ as:

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m} \tag{1}$$

$$u_{ik} = \left(\sum_{j=1}^{c} \left(\|x_k - v_i\|^2 \|x_k - v_j\|^{-2} \right)^{\frac{1}{m-1}} \right)^{-1} \tag{2}$$

Consequently, the procedure of the FCM is described as follows:

Step 1. Fix the number of clusters, the weighting exponent, the iteration limit T and the tolerance ε, and set $J_m(u, v, X) = \infty$; initialize the partition matrix;

Step 2. Update the cluster center v_i $(1 \le i \le c)$ by (1) ;

Step 3. Update the membership function u_{ik} $(1 \le i \le c,\ 1 \le k \le n)$ by (2);

Step 4. Repeat Step 2 and Step 3 until the decreasing value of $J_m(u, v, X)$ between two successive iterations is less than ε or the iterations reach T.

The objective function of the FCM can be reduced to (3) by (2), which is obtained by Bezdek in [11] as follows: $J_m(X, v) = \sum_{k=1}^{n} \left(\sum_{i=1}^{c} \|x_k - v_i\|^{\frac{-2}{m-1}} \right)^{1-m}$ (3)

It has been proved that the above algorithm converges to local minimum or saddle point of the objective function of the FCM when m>1. Let $\bar{x} = n^{-1} \sum_{k=1}^{n} x_k$, when m approaches infinite, the only solution of the FCM is $\bar{x}$ according to [1]. It is easily proved that $\bar{x}$ is a fixed point of the FCM algorithm.

As the clustering results of the FCM are greatly influenced by the weighting exponent, number of clusters, etc, it is a key issue for users to properly evaluate the clustering results of the FCM algorithm. In the literature, many methods are proposed to tackle this issue. One of the most used methods is to design an appropriate cluster validity index; Halkidi *et al* presented a well-written review of cluster validity indices in [12]. As noted in [1], many cluster validity indices like $V_{pc}(u)$ or V_{XB} have a monotone tendency with number of clusters increasing. Hence, it is always supposed that the optimal number of clusters has an upper bound $c_{\max} \le \sqrt{n}$, more details can be seen in [13]. In [1], Pal and Bezdek evaluated that the partition coefficient and entropy index, Xie-Beni Criterion, extended Xie-Beni Criterion, and the Fukuyama-Sugeno Index [14] by numerical experiments and limit analysis. They experimentally discovered that Xie-Beni Criterion provided the best response over a wide arrange of choices for number of clusters c and weighting exponent m, and the Fukuyama-Sugeno Index is not robust to both high and low values of weighting exponent m and its performance may be not stable as cluster validity index. Therefore, we use Xie-Beni index as benchmark of cluster validity index for the FCM algorithm in the following. As a matter of fact, we have another theoretical explanation of choosing a Xie-Beni index as benchmark of cluster validity index, more details will be given in Section 3.

3 A Novel Cluster Validity Index-Stability Index

As noted above, many cluster validity indices for the FCM algorithm have been proposed in the literature. However, all of them do not pay enough attention to the properties of the FCM algorithm. In this paper, we propose a novel cluster validity index based on the properties of the FCM algorithm itself. It is a reasonable assumption that the probable clustering output is the optimal clustering result of clustering algorithms if the data set complies with its clustering hypothesis. Obviously, the more stable the clustering result is, the more probable it is outputted. Therefore, we need to obtain stability criterion of clustering algorithm. Transparently, the stability criterion of a clustering algorithm depends on its optimality test. Fortunately, the optimality test of the FCM algorithm has been given in [15] or [16] as:

$$F(v) = \min_{u \in M_{fcn}} J_m(u, v, X) = \sum_{k=1}^{n} \left(\sum_{i=1}^{c} \left\| x_k - v_i \right\|^{\frac{-2}{m-1}} \right)^{1-m}$$

$$\frac{\partial^2 F(v)}{\partial v_i \partial v_j} = \frac{4m}{m-1} \sum_{k=1}^{n} u_{ik}^{\frac{m+1}{2}} u_{jk}^{\frac{m+1}{2}} \frac{(x_k - v_j)}{\left\| x_k - v_j \right\|} \frac{(x_k - v_i)^T}{\left\| x_k - v_i \right\|}$$

$$+ 2\delta_{ij} \left[\left(\sum_{k=1}^{n} u_{ik}^{m} \right) \times I_{s \times s} - \frac{2m}{m-1} \sum_{k=1}^{n} u_{ik}^{m} \frac{(x_k - v_i)(x_k - v_i)^T}{\left\| x_k - v_i \right\|^2} \right], \quad \forall 1 \le i, j \le c$$

Hessian matrix H_v of $F(v)$ can be represented by $H_v = \left(\partial^2 F(v) / \partial v_i \partial v_j \right)$. It is well known that H_v can judge whether the clustering result is stable or not, i.e., if the clustering result is a local minimum of the objective function. However, how to measure the probability the clustering result v is outputted? It is easy to conjecture that the stable degree of a clustering result is proportional to the probability the clustering result v is outputted. Therefore, we need to define an index to measure the stability of a clustering result. Since the conditional number of Hessian matrix H_v reflects the stability of Hessian matrix H_v, it can be reasonably used as an index to show the stability of a clustering result. In order to clearly visualize the experimental results in this paper, we use an ad hoc definition of the conditional number of Hessian matrix as follows: $cond(H_v) = \lambda_{\min}(H_v) / \lambda_{\max}(H_v)$, where $\lambda_{\max}(H_v)$, $\lambda_{\min}(H_v)$ are the maximum and the minimum eigenvalues of H_v, respectively.

If $1 \ge cond(H_v) > 0$, the clustering result of FCM is stable; if $cond(H_v) < 0$ or $cond(H_v) > 1$, the clustering result of FCM is unstable. Therefore, we call $cond(H_v)$ stability index. Obviously, if $1 \ge cond(H_v) > 0$, the larger $cond(H_v)$, the more stable the clustering result, therefore the more probable it is outputted. In other words, $cond(H_v)$ can measure the probability of a clustering result outputted by its corresponding algorithm. According to the above analysis, $cond(H_v)$ can be used as cluster validity index to choose the optimal number of clusters.

According to [17], undesirable solutions of clustering algorithm can be defined as follows: if $v=(v_1,v_2,\cdots,v_c)$ is an output of clustering algorithm and $\exists\ 1\le i\ne j\le c$ such that $v_i=v_j$, then it is called undesirable solutions of clustering algorithm. Noticing that Xie-Beni Criterion (in this paper, Xie-Beni Criterion and Xie-Beni index are interchangeable) is defined as:

$$V_{XB}=\left(\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^2\left\|x_k-v_i\right\|^2\right)\Big/\left(n\times\min_{1\le i\ne j\le c}\left\|v_i-v_j\right\|^2\right)$$

So, we can use Xie_Beni index to determine whether or not outputs of the FCM are undesirable solutions. But others cluster indices can not well conduct this task. For example, Fukuyama-Sugeno Index [14], partition coefficient [2], partition entropy [5], uniform data functional [6], proportion exponent [7], nonfuzziness index [8], Gunderson's separation coefficient [10]. Noticing the value of Xie_Beni index is infinite when outputs of the FCM are undesirable solutions, we use the following form instead of Xie-Beni index, which is convenient for calculation and visualization in the computer and denoted by V_{XB}^{-1}:

$$V_{XB}^{-1}=\frac{\min\limits_{1\le i\ne j\le c}\left\|v_i-v_j\right\|^2}{\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^2\left\|x_k-v_i\right\|^2}.$$

In Section 4, we verify the above conclusions by numerical experiments.

4 Numerical Experiments and Analysis

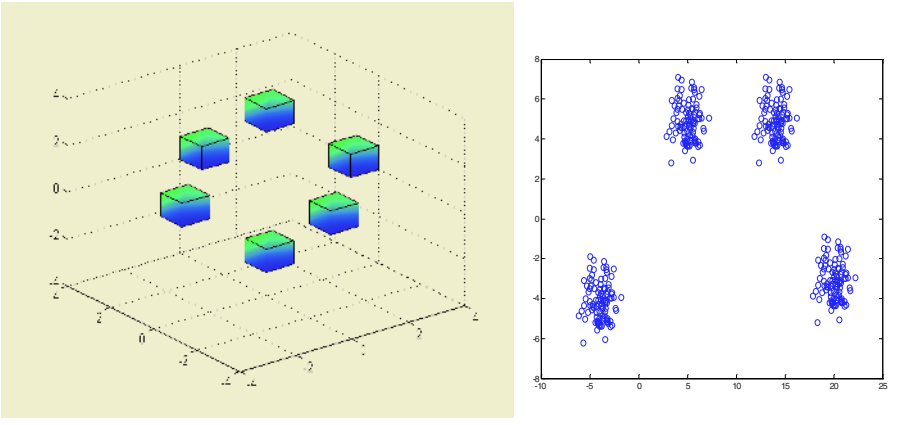
In this section, we verify the above conclusions by numerical experiments. In the following, we set the same initial partition matrix, $\varepsilon=10^{-8}$ and the maximum iteration T=200 for the FCM algorithm, and run the FCM algorithm with different weighting exponent m and number of clusters c, then calculate V_{XB}^{-1} and $cond(H_v)$ according to the clustering outputs.

The datasets used in this section are described as follows:

IRIS data: The Iris data set has 150 data points. It is divided into three groups and two of them are overlapping. Each group has 50 data points. Each point has four attributes. More details about the IRIS data are available in Anderson [18].

Cube_6. This data set is drawn as in Fig.1 (a), and consists of 6 clusters. Each cluster consists of 8 points located at 8 corners of a cube. More details about Cube6 can be seen in [16].

Data 3: Data are composed of 4 clusters as shown in Fig.1 (b). The cluster centers are as follows: $\mu_1=$ [-4, 4]; $\mu_2=$[5, 5]; $\mu_3=$[14, 5]; $\mu_4=$[20, -3]. Each cluster includes 100 points and the points in the i^{th} cluster are independently drawn from the normal distribution $N(\mu_i,I_2)$.



(a) Cube 6 (b) Data 3

Fig. 1. (a) Cube 6, (b) Data 3

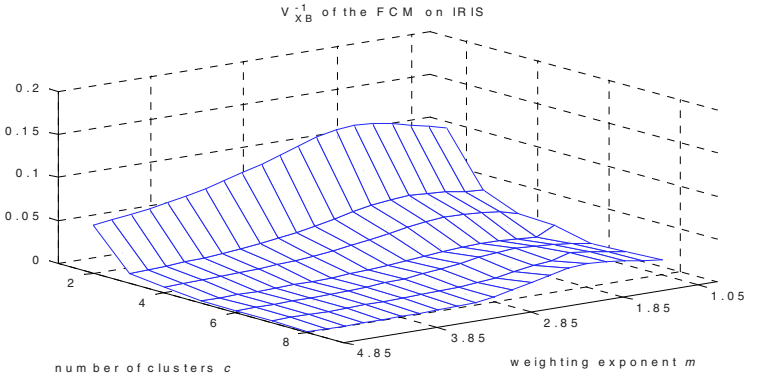


Fig. 2. V_{XB}^{-1} with varying c and m for IRIS

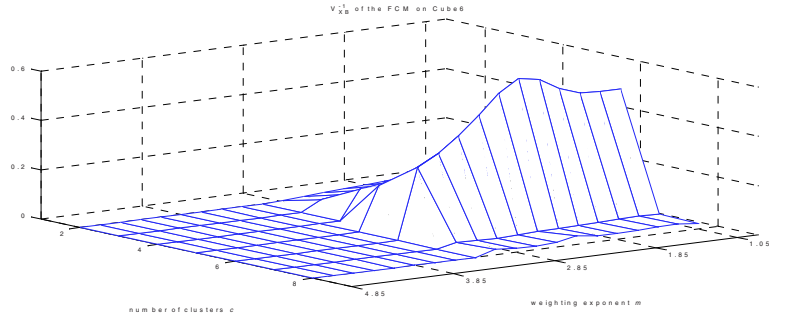


Fig. 3. V_{XB}^{-1} with varying c and m for Cube6

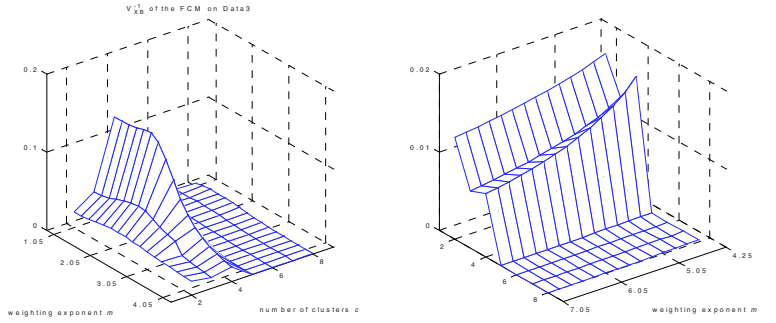
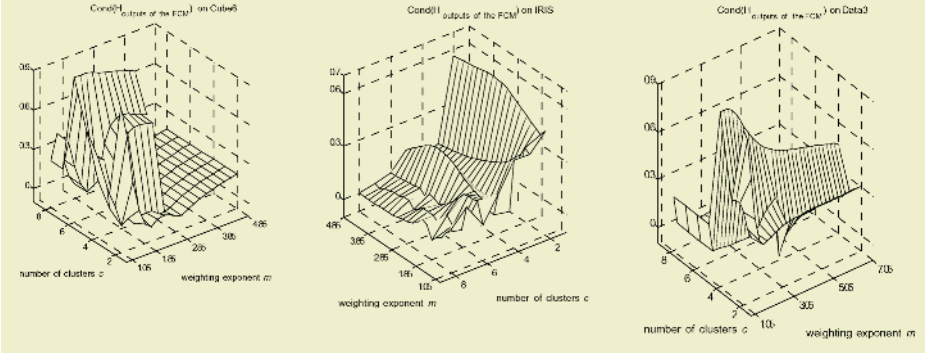


Fig. 4. V_{XB}^{-1} with varying c and m for Data 3. In the left, the weighting exponent varies from 1.05 to 4.05, V_{XB}^{-1} shows that the optimal number of clusters is 4, which is consistent with the real substructure of Data 3. However, in the right, the weighting exponent varies from 4.25 to 7.05, V_{XB}^{-1} shows that the optimal number of clusters becomes 2. According to [19], we know the FCM algorithm works well for m>1, at least in theory. It easily demonstrates that the performance of V_{XB}^{-1} sometimes heavily depends on the weighting exponent m in the FCM algorithm.



(a)Stability index with respect to c, m and Cube6

(b) Stability index with respect to c, m and IRIS

(c) Stability index with respect to c, m and data3

Fig. 5. Stability index with respect to c and m, and datasets

When m>1, Fig.5 demonstrates that the outputs of the FCM algorithm are local minimum of (3) with probability close to 1, and the performance of $cond(H_v)$ as cluster validity index is the same as Xie-Beni index. We also know that weighting exponent m plays a key role in the FCM algorithm. In [19], it is proved that if $\lambda_{\max}\left(F_{U^*_{data}}\right) < 0.5$ and $m \geq \left(1 - 2\lambda_{\max}\left(F_{U^*_{data}}\right)\right)^{-1}$, then U^*_{data} is a stable fixed point of the FCM, and if $\lambda_{\max}\left(F_{U^*_{data}}\right) \geq 0.5$, then U^*_{data} is not a stable fixed point of the FCM,

where $F_{U^*_{data}} = \left(f_{kr}^{U^*_{data}}\right)_{n\times n}$, $f_{kr}^{U^*_{data}} = \frac{1}{n}\left(\frac{x_k - \bar{x}}{\|x_k - \bar{x}\|}\right)^T\left(\frac{x_r - \bar{x}}{\|x_r - \bar{x}\|}\right)$, $U^*_{data} = \lfloor \kappa_{ij} \rfloor_{n\times n}$, $\kappa_{ij} = c^{-1}$.

It is easy to calculate that $\lambda_{\max}(F_{IRIS})$ =0.8079, $\lambda_{\max}(F_{Cube6})$ =0.3333 and $\lambda_{\max}(F_{Data\,3})$=0.7036. Therefore, we know that m<3 should hold in order to keep the FCM algorithm work well on Cube 6 according to [19]. Since $\lambda_{\max}(F_{IRIS})$=0.8079>0.5 and $\lambda_{\max}(F_{Data\,3})$=0.7036>0.5, any value of m>1 is theoretically appropriate for the FCM algorithm.

Fig. 2, 4 verify that the FCM algorithm may not outcome undesirable solutions when $\lambda_{\max}(F_{U^*_{data}}) \geq 0.5$. However, Fig.3 tells us that the FCM algorithm indeed outcomes undesirable solutions when $\lambda_{\max}(F_{U^*_{data}}) < 0.5$. Simultaneously, Fig.5 empirically proves that the outputs of the FCM algorithm are local minima, which is consistent with our intuition. As for IRIS and Cube6, the performances of $cond(H_v)$ and Xie-Beni index are the same with respect to a wide range of the weighting exponent m. As for data3, Fig.4 shows that Xie-Beni index is sensitive to high values of the weighting exponent m. However, Fig.5 shows that the performance of $cond(H_v)$ is still satisfactory with respect to a wide range of the weighting exponent m. Such facts suggest that $cond(H_v)$is more robust than Xie-Beni Criterion with respect to m as cluster validity index.

5 Conclusions and Discussions

In this paper, we propose a novel cluster validity index for the FCM algorithm, the stability index, based on the optimality test. The major contribution of this paper is that our approach is totally out of mathematical analysis of the FCM algorithm, while other previous methods out of geometrical or psychological tuition. The theoretical analysis and experimental results suggest that the stability index is valid for the FCM algorithm as a cluster validity index. Moreover, the stability index also can be used as the optimality test of the FCM algorithm.

References

1. Pal, N.R. and Bezdek, J.C. On cluster validity for the fuzzy c-means model, IEEE Trans. Fuzzy Systems, 3(3):370-379, June,1995
2. Bezdek, J.C. Pattern recognition with fuzzy objective function algorithms, Plenum Press, New York, 1981
3. Catherine A., Sugar and Gareth M.James, Finding the number of clusters in a dataset: an information-theoretic approach, Journal of the American Statistical Association, 98(463):750-763, Sept.2003

4. Tibshirani R., Walther G. & Hastie T., Estimating the number of clusters in a data set via the gap statistic, J.R.Statist.Soc.B, 63,Part 2, pp.411-423, 2001
5. Bezdek, J.C. Cluster validity with fuzzy sets, J.Cybernt., 3(3): 58-72,1974
6. Windham, M.P., Cluster validity for the fuzzy c-means clustering algorithm, IEEE Trans. PAMI, vol. PAMI-4, no.4, 357-363, July, 1982.
7. Windham, M.P., Cluster validity for fuzzy clustering algorithms, Fuzzy Sets Systems, vol.5: 177-185,1981
8. Backer, E., Jain, A.K. A Cluster performance measure based on fuzzy set decomposition, IEEE Trans. PAMI, vol.PAMI-3,no.1, Jan. 1981.
9. Xie, X.L.; Beni, G., A validity measure for fuzzy clustering, IEEE Trans. PAMI, 13 (8): 841 –847, Aug. 1991
10. Gunderson, R. Applications of fuzzy ISODATA algorithms to startracker printing systems, in Proc. 7th Triannual World IFAC Congr. 1978, pp.1319-1323
11. Bezdek, J.C. A physical interpretation of Fuzzy ISODATA, IEEE Trans. SMC, SMC-6: 387-390,1976
12. Halkidi, M., Batistakis, Y., Vazirgiannis, M. Cluster algorithms and validity measures, Proceedings of Thirteenth International Conference on Scientific and Statistical Database Management, 3 -22, 2001.
13. Yu Jian, Cheng Qiansheng, The upper bound of the optimal number of clusters in fuzzy clustering, Science in China, series F, 44(2):119-125, 2001
14. Fukuyanma, Y.and Sugeno, M. A new method of choosing the number of clusters for the fuzzy c-means method, in Proc. 5th Fuzzy Syst. Symp., 247-250(in Japanese). 1989.
15. Wei, W. and Mendel, J.M. Optimality tests for the fuzzy c-means algorithm, Pattern Recognition, vol. 27(11): 1567-1573, 1994.
16. Yu Jian, Huang Houkuan, Tian Shengfeng, An Efficient Optimality Test for the Fuzzy c-Means Algorithm, Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, vol. 1, 98-103, 2002
17. Yu Jian, General c-means clustering model and its applications, CVPR2003, v.2, 122-127, June. 2003
18. Anderson, E. The IRISes of the Gaspe Peninsula, Bull. Amer. IRIS Soc., vol.59, pp.2-5, 1935.
19. Yu Jian, Cheng Qiansheng, Huang Houkuan, Analysis of the weighting exponent in the FCM, IEEE Transactions on Systems, Man and Cybernetics-part B: Cybernetics, 34(1):634-639, Feb.2004