2005 American Control Conference
June 8-10, 2005. Portland, OR, USA

Improved Validation Index for Fuzzy Clustering

Yuangang Tang*, Fuchun Sun, *Member, IEEE* and Zengqi Sun, *Senior Member, IEEE*

***Abstract*— This paper proposes a new validation index for fuzzy clustering in order to eliminate the monotonically decreasing tendency as the number of clusters approaches to the number of data points and avoid the numerical instability of validation index when fuzzy weighting exponent increases. Limit analyses of Xie-Beni index, Kwon index and the proposed index are also considered for the convenience of contrast. Lastly, two numerical examples are presented to show the effectiveness of the proposed validation index.**

I. Introduction

CLUSTER analysis is an important mathematical tool for identifying structure presented in data set. Based on the similarity between data points, often defined by a distance measure, a number of clusters are partitioned to represent the characteristics of original data set. Fuzzy clustering, regarded as a popular method in cluster analysis, has been used extensively in pattern recognition [1], [2], image processing [3], medical diagnosis [4], fuzzy model analysis [5] etc.. As one of the best known fuzzy clustering methods the fuzzy C-means (FCM) [6] algorithm has received much attention, but some cluster validity criterions have to be required to evaluate the quality of clustering algorithm because FCM is a sort of unsupervised clustering algorithm. In order to give more accurate partitions of data, many researchers have studied this validity problem. Until now, the validation functions can be divided into two main classes according to whether the separation index is involved. One is compact index within the clusters, such as partition coefficient [7], partition entropy [8], proportion exponent [9], and the other is combined index (including fuzzy partitions and cluster centers), such as Fukuyama and Sugeno index [10], compactness and separation index [11] and Xie-Beni index [12].

The paper focuses on the Xie-Beni index. Although it can provide more reliable response over a wide range of choice for the number of clusters and fuzzy weighting exponent, Xie-Beni index has two intrinsic drawbacks: 1) validation index monotonically decreases when the number of clusters gets very large and close to data points [12], 2) there exists a very strong and unpredictable interaction between the number of clusters and fuzzy weighting exponent (numerical instability) due to its limit behavior when fuzzy weight exponent approaches to infinity. The first problem was considered by Kwon [13], who imposed an *ad hoc* punishing function to eliminate the decreasing tendency. Here, we propose an improved validation index for the FCM algorithm to overcome the above two problems with the same idea.

II. Cluster Validation Indices

Since the FCM is a popular clustering algorithm, readers interested in it may refer to [6]. Therefore, the computing formulas for alternative optimization are ignored and only cluster validation indices are considered in this paper.

Let $X=\{x_1,x_2,\cdots,x_n\}$ be a set of n data points in p-dimensional space. The $p\times n$ data matrix X has the cluster center matrix $V=[v_1,\cdots,v_c]$ ($c\in(1,n)$ is the number of clusters) and the membership matrix $U=[\mu_{ij}]_{c\times n}$, where μ_{ij} is the membership value of x_j belonging to v_i . m represents the fuzzy weighting exponent.

A. Xie-Beni index

Xie-Beni index [12] to be studied is defined as

$$V_{XB}(U,V;X)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{2}\left\|x_j-v_i\right\|^2}{n\min_{i\neq j}\left\|v_i-v_j\right\|^2} \tag{1}$$

For the first problem Xie and Beni gave a method of plotting the optimal value of $V_{XB}(U,V;X)$ for $c=2$ to $n-1$, then selecting the starting point of monotonically decreasing tendency as the maximum c to be considered. In addition, they recommended to use a punishing function imposing on the validation index, but this function was not discussed deeply. To investigate the limiting behavior of Xie-Beni index, we take two limits of the validation index when $c\to n$ and $m\to\infty$. Additional limits can be founded in [14]. Since

Manuscript submitted September 8, 2004. This work was jointly supported by the National Excellent Doctoral Dissertation Foundation (Grant No: 200041), the National Science Foundation of China (Grant No: 60474025，60334020), the National Key Project for Basic Research of China (Grant No: G2002cb312205) and Basic Research Foundation of Tsinghua University (Grant No: JC2003028).

The authors are with the Department of Computer Science and Technology, State Key Lab of Intelligent Technology and System, Tsinghua University, Beijing, 100084, P.R. China.

* Corresponding author.(phone: 0086-10-6277-7704; e-mail: tyg02@ mails.tsinghua.edu.cn)

0-7803-9098-9/05/$25.00 ©2005 AACC

$$\lim_{c\to n}\left\|x_j - v_i\right\|^2 = 0 \tag{2}$$

we obtain

$$\lim_{c\to n} V_{XB}\left(U, V; X\right) = \lim_{c\to n}\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2\left\|x_j - v_i\right\|^2}{n\min_{i\neq k}\left\|v_i - v_k\right\|^2} = 0 \tag{3}$$

On the other hand,

$$\lim_{m\to\infty}\mu_{ij} = 1/c \tag{4}$$

$$\lim_{m\to\infty} v_i = \sum\nolimits_{j=1}^{n} x_j \Big/ n = \overline{v} \tag{5}$$

where $\overline{v}$ is the fixed point of the FCM algorithm for $m > 1$, and the total scatter matrix of X

$$C_X = \sum\nolimits_{j=1}^{n}\left\|x_j - \overline{v}\right\|^2 \tag{6}$$

then we have

$$\lim_{m\to\infty} V_{XB}\left(U, V; X\right) = \lim_{m\to\infty}\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2\left\|x_j - v_i\right\|^2}{n\min_{i\neq k}\left\|v_i - v_k\right\|^2} = \frac{C_X/nc}{0} = \infty \tag{7}$$

From (3) and (7), we can see that the Xie-Beni index loses its ability to evaluate the quality of FCM as $c \to n$, and becomes unstable or unpredictable as $m \to \infty$.

B. Kwon index

In the sense of maximizing intra-class similarity and inter-class differences, Kwon [13] developed another validation index with an *ad hoc* punishing function to eliminate the monotonically decreasing tendency as the number of clusters increases. Kwon index is defined as

$$V_K\left(U, V; X\right) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2\left\|x_j - v_i\right\|^2 + \frac{1}{c}\sum_{i=1}^{c}\left\|v_i - \overline{v}\right\|^2}{\min_{i\neq k}\left\|v_i - v_k\right\|^2} \tag{8}$$

With the second term in the numerator in (8) the first problem can be effectively solved, however, the second problem can not be avoided because of its limit behavior, i.e.

$$\lim_{c\to n} V_K\left(U, V; X\right) = \lim_{c\to n}\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2\left\|x_j - v_i\right\|^2 + \frac{1}{c}\sum_{i=1}^{c}\left\|v_i - \overline{v}\right\|^2}{\min_{i\neq k}\left\|v_i - v_k\right\|^2} = \frac{C_X}{n\min_{i\neq k}\left\|v_i - v_k\right\|^2} \tag{9}$$

$$\lim_{m\to\infty} V_K\left(U, V; X\right) = \lim_{m\to\infty}\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2\left\|x_j - v_i\right\|^2 + \frac{1}{c}\sum_{i=1}^{c}\left\|v_i - \overline{v}\right\|^2}{\min_{i\neq k}\left\|v_i - v_k\right\|^2} = \frac{C_X/c + 0}{0} = \infty \tag{10}$$

Obviously, validation index $V_K\left(U, V; X\right)$ is also of numerical instability as fuzzy weighting exponent approaches infinity.

C. The proposed validation index

With the same idea of punishing function an improved validation index is defined as ($1 < c < n$)

$$V_T\left(U, V; X\right) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2\left\|x_j - v_i\right\|^2 + \frac{1}{c(c-1)}\sum_{i=1}^{c}\sum_{k=1,k\neq i}^{c}\left\|v_i - v_k\right\|^2}{\min_{i\neq k}\left\|v_i - v_k\right\|^2 + 1/c} \tag{11}$$

The second term in the numerator in (11) is an *ad hoc* punishing function (average distance between cluster centers) applied to eliminate the decreasing tendency as $c \to n$, moreover, the second term in denominator in (11) is also a punishing function used to strengthen the numerical stability as $m \to \infty$. Then it can be obtained that

$$\lim_{c\to n} V_T\left(U, V; X\right) = \lim_{c\to n}\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2\left\|x_j - v_i\right\|^2 + \frac{1}{c(c-1)}\sum_{i=1}^{c}\sum_{k=1,k\neq i}^{c}\left\|v_i - v_k\right\|^2}{\min_{i\neq k}\left\|v_i - v_k\right\|^2 + 1/c} = \frac{\sum_{i=1}^{n}\sum_{k=1,k\neq i}^{n}\left\|x_i - x_k\right\|^2}{n(n-1)\min_{i\neq k}\left\|x_i - x_k\right\|^2 + (n-1)} \tag{12}$$

When $m \to \infty$, we have

$$\lim_{m\to\infty}\left\|v_i - v_k\right\| = 0 \tag{13}$$

Without the punishing function in the denominator in (11), the proposed index will go to infinity due to (5) and (13). To avoid this, based on (4) we select the punishing function as $1/c$ such that the limit of (11) converges to an inherent metric of X, i.e. (6). And then, we obtain

$$\lim_{m\to\infty} V_T\left(U, V; X\right) = \lim_{m\to\infty}\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2\left\|x_j - v_i\right\|^2 + \frac{1}{c(c-1)}\sum_{i=1}^{c}\sum_{k=1,k\neq i}^{c}\left\|v_i - v_k\right\|^2}{\min_{i\neq k}\left\|v_i - v_k\right\|^2 + 1/c} = C_X \tag{14}$$

From the above two limits we can conclude that the proposed validation index can keep the ability to discriminate between various values of clusters as $c \to n$, and also assure the numerical stability of validation index as $m \to \infty$. That is to say, the proposed validation index can provide better response over a wide range of choices for the number of clusters and fuzzy weighting exponent than Xie-Beni index and Kwon index.

Remark 1: Based on the above comparisons, some facts should be emphasized:

1) All the three indices have considered the compactness and separation of fuzzy partition.

2) Xie-Beni index has no punishing functions, Kwon index has an *ad hoc* punishing function in its numerator, and the proposed index has an *ad hoc* punishing function in its numerator and denominator respectively. With these modifications the limit behavior of the indices can be improved effectively. More importantly, the limits of the proposed index depend on the inherent characteristics of the given data set X, such as n, C_X.

III. NUMERICAL EXAMPLE

To compare the performance of the proposed validation index with Xie-Beni index and Kwon index, an evaluation study is carried out. This study involves two numerical examples. For sake of showing the differences in the three validation index some common parameters of FCM are fixed: terminating criterion $\varepsilon = 0.001$, $\|\bullet\|$ is the Euclidean norm, and the initial centroids are randomly chosen as distinct points in data set for different number of clusters. For a particular c the initial values are the same. In particular, numerical instability should be emphasized because it could lead to many different results with different initial values. Lots of experiments have been done with different initializations which are not reported here, and those with smaller exponent are very similar to the ones given in the following tables. Minimizing the three indices may give different optimal values. In the tables the bold values correspond to the optimal values of c chosen by each index and the italic values denote the unstable or unpredictable values of each index.

Example 1: In this example the famous butterfly data set [6] is employed, which has $c^* = 2$ as the number of preferred clusters. Table 1, 2 and 3 report the index values for Xie-Beni index, Kwon index and the index proposed in this paper, respectively.

From table 1 we can see that Xie-Beni index loses its ability to validate (U, V) pairs from FCM as $c \to n$ and is numerically unstable as m increases. Kwon index in table 2 is also numerically unstable due to its limit behavior, such as the point $(m = 14, c = 13)$ and the point $(m = 16, c = 10)$. The proposed index shown in table 3 gives correct clusters in a wider range of c and keeps strong numerical stability when m increases.

Example 2: In this example the well-known IRIS data set [15] is considered. Although coming from three physical clusters each with 50 points, IRIS has two well-separated geometrical structure. Since clusters are represented by mathematical properties of the data set, we regard $c^* = 2$ as the optimal number of the cluster centers for IRIS. Table 4, 5 and 6 are the index values for Xie-Beni index, Kwon index and the proposed index, respectively.

TABLE I XIE-BENI INDEX FOR BUTTERFLY DATA SET

m \ c	2	3	4	5	6	7	8	9	10	11	12	13	14
2	0.0954	0.2097	0.3994	0.1061	0.0946	0.0649	0.1123	0.1558	0.1130	0.0995	0.0682	0.0459	**0.0207**
4	0.1611	1.21	0.7608	1.0538	0.8024	0.6500	0.5349	0.4340	0.3534	0.2796	0.2077	0.1315	**0.0670**
6	0.2088	0.5122	1.5148	1.1919	0.9686	0.7588	0.6118	0.5098	0.3906	0.3280	0.2444	0.1530	**0.0789**
8	0.2362	2.2970	1.6431	1.3698	1.0397	0.8202	0.6485	0.5494	0.4240	0.3228	0.2631	0.1639	**0.0851**
10	0.2539	0.5963	1.7131	1.4196	1.0964	0.8606	0.6823	0.5733	0.4441	0.3157	0.2743	0.1704	**0.0319**
12	0.2662	0.6141	1.7596	68.622	1.1346	0.8928	0.7049	0.5891	0.4575	0.32485	0.2817	0.1747	**0.0328**
14	**0.2751**	2.5933	1.7929	1.4038	1.1508	0.91048	0.7210	0.5746	0.4670	0.3661	0.2781	*8.67e+5*	56.292
16	**0.2819**	2.6366	1.818	1.4226	1.1698	0.9243	0.7331	0.5909	*6.32e+10*	8.1132	14.632	*9797.7*	23

TABLE II KWON INDEX FOR BUTTERFLY DATA SET

m \ c	2	3	4	5	6	7	8	9	10	11	12	13	14
2	**1.6803**	3.8254	8.5415	2.7356	3.1882	2.9809	6.0524	10.025	7.8732	8.6065	8.7904	8.2929	7.0118
4	**2.667**	20.371	13.342	21.227	16.131	13.79	11.281	10.303	9.6099	8.5612	8.2019	7.6656	7.2198
6	**3.3819**	8.3503	25.217	19.883	17.198	14.811	12.802	11.425	10.56	9.2833	8.7495	7.9877	7.3982
8	**3.7937**	37.394	27.147	25.948	21.095	15.731	14.353	12.019	11.06	9.9322	9.0296	8.1503	7.4906
10	**4.0589**	9.6105	28.196	26.494	21.946	15.908	14.86	12.377	11.361	10.553	9.1975	8.2478	7.6208
12	**4.2424**	9.8774	28.894	1160.5	22.518	16.249	15.199	12.615	11.562	10.691	9.3092	8.3131	7.6352
14	**4.3764**	41.897	29.394	23.857	22.096	17.8	15.44	13.286	11.705	10.946	9.5888	*2.47e+07*	*1822*
16	**4.4786**	42.548	29.77	24.14	22.38	18.007	15.621	13.975	*1.46e+12*	173.03	352.96	*1.80e+05*	584.07

TABLE III THE PROPOSED INDEX FOR BUTTERFLY DATA SET

m \ c	2	3	4	5	6	7	8	9	10	11	12	13	14
2	**2.366**	4.8967	11.309	4.2787	5.3841	5.4632	10.893	17.595	14.071	15.317	16.373	15.605	13.658
4	**3.3162**	20.148	14.48	24.214	18.678	16.757	13.626	13.486	13.423	12.652	12.938	13.286	13.299
6	**4.0056**	8.9619	23.555	19.075	17.715	16.877	15.491	14.483	14.721	13.31	13.44	13.584	13.464
8	**4.4037**	32.396	25.054	28.29	24.624	17.681	17.885	15.017	15.176	14.671	13.699	13.735	13.55
10	**4.6605**	10.103	25.891	28.578	25.353	17.378	18.336	15.339	15.45	16.007	13.854	13.825	14.805
12	**4.8384**	10.348	26.448	126.91	25.844	17.385	18.637	15.553	15.632	16.133	13.957	13.886	14.818
14	**4.9683**	35.675	26.848	23.214	24.682	20.2	18.852	17.207	15.763	15.6	14.634	203.73	230.84
16	**5.0675**	36.162	27.149	23.45	24.925	20.381	19.013	18.077	175.29	102.79	145.07	351.24	196.71

In the above three tables only the values for $m = 2,10,11,12,13$ are enumerated due to the limitation of space, and the numerical characteristics are similar when m increases. It is very clear from table 4, 5 and 6 that many values for Xie-Beni index and Kwon index are unpredictable for large values of m, however, the proposed validation index shows better performance than the two indices for its preferable limit behavior.

Remark 2: As an important parameter fuzzy weighting exponent can affect fuzzy memberships and cluster centers, which can be seen from (4) and (5). Readers interested in how to select a proper m may refer to [16]. Here, our numerical results reveal that even without optimal value for m the proposed index can give a reliable fuzzy partition over a wide range of choice for the number of clusters and fuzzy weighting exponent.

TABLE IV XIE-BENI INDEX FOR IRIS DATA SET

m \ c	2	3	4	5	6	7	8	9	10	11
2	**0.0542**	0.1369	0.1953	0.2277	0.3109	0.3744	0.4897	0.3672	0.3236	0.3819
10	**0.1947**	1.5379	1.8751	*3.48e+7*	*1.35e+5*	27805	210.91	*1.89e+6*	*1.31e+10*	11521
11	**0.2025**	1.6622	2.0077	*2.18e+8*	*5.13e+6*	3647.8	308.23	*5.98e+5*	*1.37e+8*	*6.0e+6*
12	**0.2091**	1.7738	2.1222	*2.72e+7*	*1.65e+10*	1920.7	344.02	*2.10e+5*	*1.28e+5*	1498
13	**0.2148**	1.8692	2.2156	*1.32e+8*	*1.55e+8*	3514.9	351.8	*1.04e+5*	11612	1912.1

TABLE V KWON INDEX FOR IRIS DATA SET

m \ c	2	3	4	5	6	7	8	9	10	11
2	**8.3943**	21.958	32.066	38.788	55.973	68.911	86.539	71.384	64.99	74.304
10	**29.469**	233.62	285.43	*5.34e+9*	*2.07e+7*	*4.29e+6*	32580	*2.95e+8*	*2.04e+12*	*1.8e+6*
11	**30.64**	252.38	305.44	*3.34e+10*	*7.89e+8*	*5.62e+5*	47563	*9.30e+7*	*2.14e+10*	*9.38e+8*
12	**31.634**	269.21	322.73	*4.16e+9*	*2.54e+12*	*2.96e+5*	53032	*3.26e+7*	*2.00e+7*	*2.34e+5*
13	**32.486**	283.61	336.8	*2.03e+10*	*2.38e+10*	*5.41e+5*	54191	*1.61e+7*	*1.81e+6*	*2.98e+5*

TABLE VI THE PROPOSED INDEX FOR IRIS DATA SET

m \ c	2	3	4	5	6	7	8	9	10	11
2	**0.0542**	0.1369	0.1953	0.2277	0.3109	0.3744	0.4897	0.3672	0.3236	0.3819
10	**29.254**	188.25	217.71	920.74	899.87	888.75	852.78	917.05	905.98	901.08
11	**30.384**	201.33	230.54	953.16	931.08	915.24	887.32	950.62	935.56	929.38
12	**31.345**	212.88	241.47	980.78	956.86	933.47	907.71	976.22	964.79	953.2
13	**32.167**	222.68	250.34	1004.2	978.45	953.49	924.7	997.7	984.69	973.36

IV. CONCLUSION

In this paper, an improved validation index based on Xie-Beni index is proposed in order to eliminate the monotonically decreasing tendency as the number of clusters approaches to the number of data points and avoid the numerical instability of cluster validation index when fuzzy weighting exponent increases. Moreover, limit analysis of the three indices are carried out because it is not always possible to regard the number of clusters and the value of fuzzy exponent as *a prior* knowledge. In the light of limit behavior we may check other validation indices or construct better indices for fuzzy clustering.

REFERENCES

[1] Baraldi, A., Blonda, P, "A survey of fuzzy clustering algorithms for pattern recognition. I.," *IEEE Trans. Systems, Man and Cybernetics, Part B*, vol. 29, no. 6, pp.778 - 785, Dec. 1999.

[2] Baraldi, A., Blonda, P, "A survey of fuzzy clustering algorithms for pattern recognition. II.," *IEEE Trans. Systems, Man and Cybernetics, Part B*, vol. 29, no. 6, pp.786 - 801, Dec. 1999

[3] Shu-Hung Leung, Shi-Lin Wang, Wing-Hong Lau, "Lip image segmentation using fuzzy clustering incorporating an elliptic shape function," *IEEE Trans. Image Processing,* vol. 13, no. 1, pp. 51 - 62, Jan. 2004.

[4] Amini, L., Soltanian-Zadeh, H., Lucas, C., Gity, M., "Automatic segmentation of thalamus from brain MRI integrating fuzzy clustering and dynamic contours," *IEEE Trans. Biomedical Engineering,* vol. 51, no. 5, pp. 800-811, May 2004.

[5] Nascimento, S., Mirkin, B., Moura-Pires, F., "Modeling proportional membership in fuzzy clustering," *IEEE Trans. on, Fuzzy Systems,* vol. 11, no. 2, pp.173 - 186, April 2003.

[6] James C.Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[7] J.C. Bezdek, "Cluster validity with fuzzy sets," *J. Cybernet*, vol.3, pp. 58 - 73, 1973.

[8] J.C.Bezdek, "Mathematical models for systematics and taxonomy," in Proc*. 8th Int. Conf. Numerical Taxonomy*, G. Estabrook, Ed., Freeman, San Franscisco, CA, pp. 143 – 166, 1975.

[9] M.P. Windham, "Cluster validity for fuzzy clustering algorithms," *Fuzzy Sets and Syst.,* vol. 5, pp177 - 185, 1981.

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

[11] B.L.M.R. Rezae and J. Reiber, "A new cluster validity index for the fuzzy c-means," *Pattern Recognition Letters,* vol. 19, pp. 237 - 246, 1998.

[12] X. Xie and G. Beni, "A Validity Measure for Fuzzy Clustering," *IEEE Trans. Pattern Analysis and Machine* Intelligence *(PAMI),* vol. 13, no. 8, pp.841 - 847, 1991.

[13] Kwon, S.H., "Cluster validity index for fuzzy clustering," *Electronics Letters,* Vol. 34, Issue: 22, pp.2176 - 2177, 1998.

[14] Pal, N.R.; Bezdek, J.C, "On cluster validity for the fuzzy c-means model," *IEEE Trans. Syst.,* vol. 3, no.3, pp.370 - 379, 1995.

[15] Bezdek, J.C.; Keller, J.M.; Krishnapuram, R.; Kuncheva, L.I.; Pal, N.R., "Will the real iris data please stand up?" *IEEE Trans. Fuzzy Systems,* vol. 7, no. 3, pp.368 - 369, 1999

[16] Jian Yu, Qiansheng Cheng, Houkuan Huang, "Analysis of the weighting exponent in the FCM," *IEEE Trans. Systems, Man and Cybernetics, Part B,* vol. 34, no. 1, pp. 634 - 639, Feb. 2004.