A Validity Index for Fuzzy and Possibilistic C-means Algorithm

Chunhui Zhang
Beihang University
prospring@163.com

Yiming Zhou
Beihang University
zhouyiming@buaa.edu.cn

Trevor Martin
University of Bristol
trevor.martin@bristol.ac.uk

Abstract

This paper proposes a novel validity index for *fuzzy-possibilistic c-means(FPCM)* algorithm, it combines *extended partition entropy* and inter class similarity which is calculated from the fuzzy set point of view. The proposed index only requires the *membership* matrix and *possibilistic(typicality)* matrix, and is free from heavy distance computing. We also extend Xie-Beni index and Kwon index to evaluate FPCM. Experiments are done to compare the three indices and the results show its effectiveness.
Keywords: Fuzzy clustering, Validity index

1 Introduction

Clustering algorithms are unsupervised learning methods. Their outputs are sensitive to predefined parameters. The same algorithm can produce different outcomes with different parameters. In the literature there are many studies on how to choose optimal parameters in clustering algorithms[8][9]. Such problems are called cluster validity problems.

In this paper, we investigate validity indices suitable for *fuzzy-possibilistic c-means* and use them to find the optimal number of clusters for a data set.

In general, there are three kinds of validity indices for fuzzy clustering. The first kind of indices involve only the membership values and are based on the assumption that the outputs are better if they are closer to a crisp partition. These indices include PC(partition coefficient)[4] , PE(partition entropy)[3], uniform data functional[14], proportion exponent[13], nonfuzziness index [2],etc. The second class of methods take into account geometrical properties of the data set, for instance, Xie-Beni(XB) index[15],Fukuyama-Sugeno index[6], and Kwon index[12]. They involve both the *membership matrix* and the data set itself. Jian Yu and Cui-Xia Li developed a cluster validity index for the Fuzzy C-Means algorithm, based on the optimality test, called stability index for FCM(Fuzzy c-means) [7]. It is different from the above two kinds of indices because it pays more attention to the clustering algorithm and relates cluster validity to stability of clustering algorithms.

In this paper, we first extend *XB(Xie-Beni)* index and *Kwon* index to determine the optimal cluster number for *FPCM*[10]. Then we propose a novel and simple validity index for *FPCM*, which uses only the calculated *membership matrix* and *possibilistic matrix* based mainly on fuzzy set theory.

The remainder of this paper is arranged as follows. Section 2 gives a brief introduction to the *FPCM* algorithm; Section 3 presents the extended *XB* and *Kwon* index; Section 4 gives a detailed description of the proposed *EPESIM* validity index; Section 5 includes all the experiments done to compare and analyze the three indices; Section 6 gives conclusions

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 877–882
Torremolinos (Málaga), June 22–27, 2008

and discusses future work.

2 FPCM Algorithm

FPCM[10] algorithm was proposed by N.R.Pal, K.Pal, and J.C.Bezdek, it includes both *possibility(typicality)* and *membership* values.*FPCM* model can be seen as the following optimization problem:

$$\underbrace{\min}_{(U,T,V)} \{J_{m,\eta}(U,T,V;X)\} = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik}^m + t_{ik}^\eta)D_{ikA}^2 \tag{1}$$

subject to the constraints $m > 1$,$\eta > 1$,$0 \leq u_{ik}$,$t_{ik} \leq 1$, $D_{ikA} = \|x_k - v_i\|_A$, and

$$\sum_{i=1}^{c} u_{ik} = 1 \forall k, i.e., U \in M_{fcn} \tag{2}$$

and

$$\sum_{k=1}^{n} t_{ik} = 1 \forall i, i.e., T^t \in M_{fnc}. \tag{3}$$

Where U is *membership* matrix, T is *possibilistic* matrix, and V is the resultant cluster centers, c and n are cluster number and data point number respectively. The first order necessary conditions for extreme of $J_{m,\eta}$ are: If $D_{ikA} = \|x_k - v_i\|_A > 0$ for all i and $k, m, \eta > 1$, and X contains at least c distinct data points,then $(U, T^t, V) \in M_{fcn} \times M_{fcn} \times \mathbb{R}^p$ may minimize $J_{m,\eta}$ only if

$$u_{ik} = (\sum_{j=1}^{c}(\frac{D_{ikA}}{D_{jkA}})^{2/(m-1)})^{-1}$$
$$1 \leq i \leq c; 1 \leq k \leq n \tag{4}$$

$$t_{ik} = (\sum_{j=1}^{n}(\frac{D_{ikA}}{D_{ijA}})^{2/(\eta-1)})^{-1}$$
$$1 \leq i \leq c; 1 \leq k \leq n \tag{5}$$

and

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik}^m + t_{ik}^\eta)x_k}{\sum_{k=1}^{n}(u_{ik}^m + t_{ik}^\eta)}, 1 \leq i \leq c. \tag{6}$$

The above equations show that *membership* u_{ik} is affected by all c cluster centers, while *possibility* t_{ik} is affected only by the *i-th* cluster center c_i. The possibilistic term distributes the t_{ik} with respect to all n data points, but not with respect to all c clusters. So, *membership* can be called *relative typicality*,it measures the degree to which a point belongs to one cluster relative to other clusters and is used to crisply label a data point. And *possibility* can be viewed as *absolute typicality*, it measures the degree to which a point belongs to one cluster relative to all other data points, it can reduce the effect of outliers. Combining both *membership* and *possibility* can lead to better clustering result.

3 Extended Xie-Beni and Kwon Index

Present fuzzy validity indices are all for *fuzzy c-means algorithm* and alike. The original indices for *FCM* can not be used directly to evaluate *FPCM*,because the FPCM algorithm generates both membership and possibility for all points to all clusters, while the original indices only consider the membership matrix, which is not sufficient. So it's necessary to find new validity indices for it.

3.1 Original indices for FCM

Xie-Beni index is defined as

$$S = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{i,k}^2 \|V_i - X_k\|^2}{n \min_{i \neq j} \|V_i - V_j\|^2} \tag{7}$$

A smaller S means all the clusters are overall compact and separate in a partition.
The *Kwon index* is based on *Xie-Beni index*, and can be used to eliminate the monotonically decreasing tendency when the number of clusters becomes very large and close to the number of data points. It is defined as

$$v_K(U,V;X) = (\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^2 \|x_k - v_i\|^2 + \frac{1}{c}\sum_{i=1}^{c}\|v_i - \overline{X}\|^2) / \min_{i \neq j}(\|v_i - v_j\|^2) \tag{8}$$

where $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} x_j$. The second item of the numerator approaches a positive constant $\frac{1}{n}\sum_{i=1}^{n}\|v_i - \overline{X}\|^2$ when c approaches n.

3.2 Extended Indices for FPCM

We extend *XB index* to

$$S' = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{i,k}^2 + t_{i,k}^2)\|V_i - X_k\|^2}{n \min_{i\neq j}\|V_i - V_j\|^2} \quad (9)$$

which takes into account both *membership* and *possibility(typicality)*, and can be used to validate the partition got by *FPCM* algorithm. We call it *EXB(Extended XB Index)*. Similarly, the *Kwon index* can also be extended to

$$V'(U,T,V;X) = (\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik}^m + t_{ik}^n)\|x_k - v_i\|^2 + \frac{1}{c}\sum_{i=1}^{c}\|v_i - \overline{X}\|^2)/\min_{i\neq j}(\|v_i - v_j\|^2) \quad (10)$$

Which we call *EKwon*.

4 A Novel Cluster Validity Index from Fuzzy Set Point of View

Many validity indices take into account two aspects of a partition, cluster *compactness* and *separation*. For *XB index* and *Kwon Index*, cluster compactness is modeled by the overall distance from all the data points to each cluster center. And cluster separation is calculated by the minimum distance between cluster centers. In this paper, we try to find a easier way to model the two aspects of a fuzzy partition. We assume that for a given data set, the *compactness* and *separation* of a partition can be obtained from the resulted *membership matrix U* and *possibility matrix T*. It's the starting point which motivates our index.

4.1 Inter-class Similarity

Each fuzzy cluster of a partition can be seen as a fuzzy set, and the whole data set is the universe for them. Then the separation between clusters can be modeled by the similarity between all these fuzzy sets. Less similarity means better separation.
We use the following equation to calculate the similarity of two fuzzy sets[1]

$$S(A,B) = \frac{\|A \cap B\|}{\|A \cup B\|} \quad (11)$$

where $\|.\|$ is the cardinality of fuzzy set, A and B are fuzzy sets. In this paper, we use sigma count to compute the cardinality of a fuzzy set. There are many other methods to calculate the similarity between fuzzy sets besides equation11. We will try to investigate on alternative similarity measures in the future.
The similarity of two fuzzy clusters depends on both their *membership* vector and *possibility(typicality)* vector. So for each pair of clusters, we need to calculate two aspects of similarity.
The *mean inter-class similarity* of *membership* matrix is

$$\overline{S}(U) = \frac{1}{c\times(c-1)/2}\sum_{i\neq j}S(U(i),U(j)) = \frac{1}{c\times(c-1)/2}\sum_{i\neq j}\frac{\sum_{k=1}^{n}\min(u_{ik},u_{jk})}{\sum_{k=1}^{n}\max(u_{ik},u_{jk})} \quad (12)$$

$U(i)$ and $U(j)$ are different rows of U. The value for $\overline{S}(U)$ is between zero and one.
And the *mean inter-cluster similarity* of *possibility* matrix is

$$\overline{S}(T) = \frac{1}{c\times(c-1)/2}\sum_{i\neq j}S(T(i),T(j)) = \frac{1}{c\times(c-1)/2}\sum_{i\neq j}\frac{\sum_{k=1}^{n}\min(t_{ik},t_{jk})}{\sum_{k=1}^{n}\max(t_{ik},t_{jk})} \quad (13)$$

Then the overall inter-class similarity is to combine them together

$$SIM(U,T) = \alpha_1 \times \overline{S}(U) + \beta_1 \times \overline{S}(T) \quad (14)$$

where α_1 and β_1 are predefined constants and should satisfy $\alpha_1 + \beta_1 = 1$.

4.2 Extended Partition Entropy

From intuitive point of view, for each compact cluster, data points distribute densely near the center of the cluster. While for less compact cluster, data points belonging to it distribute sparsely. For fuzzy clustering, a com-

pact cluster means there are some points belonging to it with high membership and typicality. This implies the membership and typicality vary a lot among all the data points to the cluster. Compact clusters can give more information than less compact clusters. We use the idea of partition entropy to model the *compactness* of fuzzy partition.
The original PE[3] proposed by Bezdek is defined as

$$PE = -\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}\log(u_{ik}) \qquad (15)$$

We apply it to both *membership matrix* and *possibilistic matrix*.
Partition entropy for U is

$$PE(U) = -\frac{1}{n\times\log(c)}\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}\log(u_{ik}) \qquad (16)$$

The added $\log(c)$ is used to normalize it, which can eliminate the original *PE's* preference to smaller cluster number. The *Partition Entropy* for T is similar to *U's*, it is defined as

$$PE(T) = -\frac{1}{c\times\log(n)}\sum_{i=1}^{c}\sum_{k=1}^{n} t_{ik}\log t_{ik} \qquad (17)$$

We combine them with weights, the resulted equation is called *EPE(Extended Partition Entropy)*

$$EPE(U,T) = \alpha_2\times PE(U)+\beta_2\times PE(T) \qquad (18)$$

where α_2 and β_2 are similar to those in formula(14).

At last, we integrate *EPE* and *SIM* into one equation, then the whole validity index we proposed is:

$$PESIM = \alpha_3\times EPE(U,T)+\beta_3\times SIM(U,T) \qquad (19)$$

where α_3 and β_3 are also weighting parameters for *EPE* and *SIM*.

5 Experiments

This section shows the experiments comparing *EXB*,*EKwon*, and our proposed *EPESIM*.

All the tests are done on DELL INSPIRON 640m laptop, with a duo-core CPU,1.6GHz for each core and RAM of 1G. The operating system is Ubuntu 7.04, and Matlab R2007a is the programming software. In *EPESIM*,*EPE* and *mean inter-class similarity* are weighted by $0.7(\alpha_3)$ and $0.3(\beta_3)$ respectively, and other parameters,α_1,β_1,α_2,β_2 are all chosen to be 0.5. For *FPCM* model, m and η are set to be 2.
There are four testing data sets, they are IRIS,X1,X2,X3. The last three data sets are generated according to different goals. All index values are the mean values of ten observations. For each index, the cluster number corresponding to the minimum index value is the optimal cluster number that index gets.
X1 includes three well separated clusters. And all the validity indices get correct cluster number 3 for it.

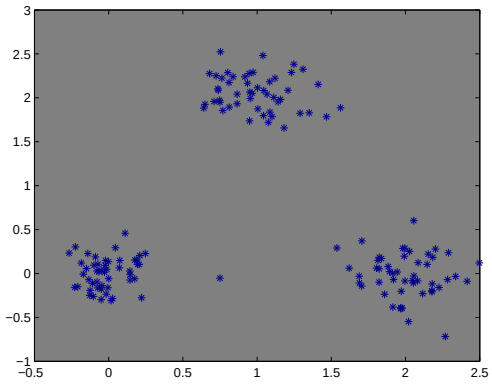

Figure 1: Dataset X1

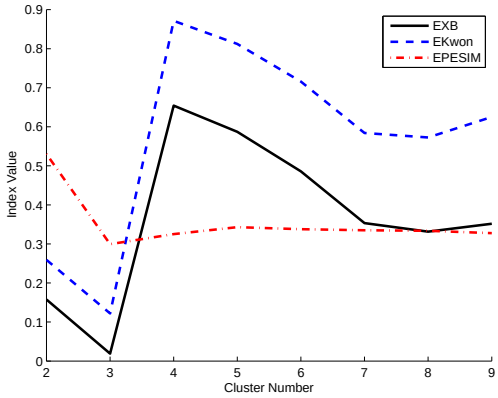

Figure 2: Result for X1

Then we use IRIS data set in the UCI database to compare the three indices. IRIS contains 150 4-dimensional data points belonging to three different clusters. And two of the clusters overlap. In Table1, the values marked by asterisks are the lowest values for the three indices and the corresponding cluster numbers are the optimal cluster numbers those indices choose. Table1 shows that,*EPESIM* get correct cluster number 3 for IRIS. And the other two indices both choose 2 as its optimal cluster number.

X2 includes four clusters, two of them are

Table 1: Result for IRIS

C	EXB	EKwon	EPESIM
2	0.0557*	0.0575*	0.4006
3	0.1133	0.1216	0.3986*
4	0.3553	0.3916	0.4397
5	0.3506	0.3992	0.4527
6	0.3580	0.4083	0.4618
7	0.4138	0.4908	0.4657
8	0.3690	0.4532	0.4711
9	0.5050	0.6431	0.4688

quite close and the other two are relative detached, this data set is used to compare the indices' behavior under situations of non evenly distributed clusters.

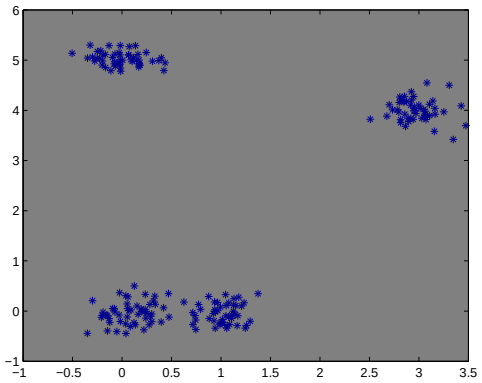

Figure 3: Dataset X2

Table2 gives the exact result for X2, *EPESIM* gets 4, while *EXB* and EKwon get 2.

X3 is the most complicated data set, it comprises five normal distributed clusters with different cluster center and deviation. The

Table 2: Result for X2

C	EXB	EKwon	EPESIM
2	0.0620*	0.0633*	0.4444
3	0.1280	0.1339	0.2687
4	0.7707	0.9424	0.2572*
5	1.2895	1.6898	0.2966
6	0.6951	1.0797	0.2971
7	0.4849	0.8761	0.3042
8	0.4467	0.8810	0.3186
9	0.4183	0.8667	0.3051

two bottom clusters are adjacent, the top three clusters are more complex, two of them overlap a little, and the other one is a little sparse.

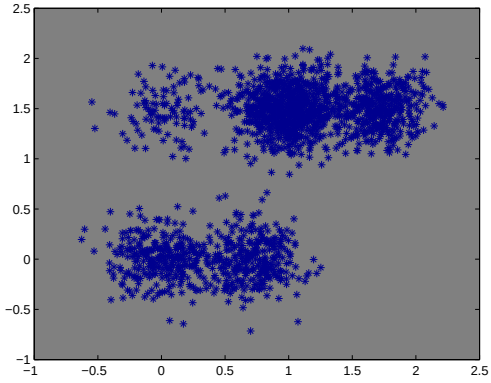

Figure 4: Dataset X3

EPESIM also gets the right answer for X3. The experiments show that for well separated clusters, all the indices can get right answer. And for non evenly distributed and more complicated situations, *EPESIM* is better. *EXB* and *EKwon* both prefer smaller cluster number than the correct one.

6 Conclusion and Future Work

This paper proposes a cluster validity index based on the combination of extended partition entropy and inter-class similarity which is calculated from the point of fuzzy logic. Experiments show that the proposed index gets prominent results under various kind of situations. And the most important advantage of the proposed *EPESIM* is that, it only uses the

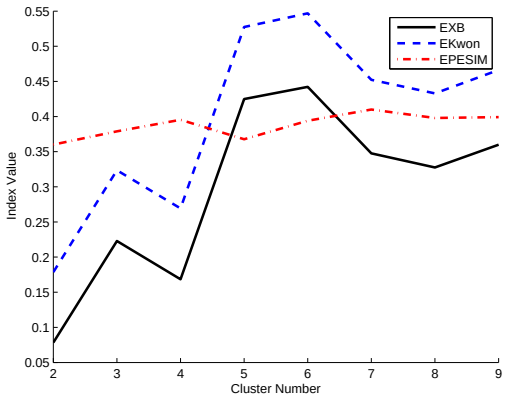

Figure 5: Result for X3

resultant *membership matrix* and *possibilistic matrix*, and doesn't involve the time consuming distance computing among the cluster centers and between all the data points and the cluster centers. Experiments show that this simple method doesn't lead to worse result.

In the future, we will investigate on alternative methods to calculate similarity between two fuzzy sets, and the behavior of the proposed index under more situations. We will also try to find theoretical instructions for selecting weighting parameters.

References

[1] D.Dubois and H.Prade. Fuzzy Sets and Systems-Theory and Applications. New York: Academic Press, 1980. *Information and Control*, volume 20, pages 301-312, 1972

[2] Backer E,Jain A K. A Cluster performance measure based on fuzzy set decomposition. *IEEE Trans.PAMI*, volume 3(1), Jan.1981.

[3] Bezdek J C. Cluster va lidity with fuzzy sets. *J.Cybernt.*, volume 3(3), pages 58-72,1974.

[4] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, NewYork, 1981.

[5] D.Dubois and H.Prade. Fuzzy Sets and Systems, Theory and Applications. Academic Press, NewYork, 1980.

[6] Fukuyanma Y. and Sugeno M. A new method of choosing the number of clusters for thefuzzy c-means method. In *Proc.5th Fuzzy Syst.Symp.*, pages 247-250(in Japanese), 1989.

[7] Jian Yu, Cui-Xia Li. Novel Cluster Validity Index for FCM Algorithm. *J.Comput.Sci.& Technol.*, volume 21, No.1, pages 137-140, Jan.2006.

[8] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis. Cluster Validity Methods:PartI. *SIGMOD Rec.*, volume 31, No.2, pages 40-45, 2002.

[9] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis. Cluster Validity Methods:PartII. *SIGMOD Rec.*, volume 31, No.3, pages 19-27, 2002.

[10] N.R.Pal, and J.C.Bezdek. A mixed c-means clustering model. In *IEEE Int.Conf.Fuzzy Systems*, pages 11-21, Spain, 1997,.

[11] N.R.Pal, K.Pal, J.M.Keller and J.C.Bezdek. A Possibilistic Fuzzy c-Means Clustering Algorithm. *IEEE TRANS.ON FUZZY SYSTEMS*, volume 13, NO.4, pages 517-530, AUGUST 2005.

[12] S.H.Kwon. Cluster validity index for fuzzy clustering. *ELETRONICS LETTERS*, volume 34(22), pages 2176-2177, 1998.

[13] Windham M P. Cluster validity for fuzzy clustering algorithms. *Fuzzy Sets Systems*, volume 5, pages 177-185,1981.

[14] Windham M P. Cluster validity for the fuzzy c-means clustering algorithm. *IEEE Trans.PAMI*, volume 4(4), pages 357-363, July 1982.

[15] Xie X L, Beni G. A validity validity measure for fuzzy clustering. *IEEE Trans.PAMI*, volume 13(8), pages 841-847,Aug.1991.