New Fuzzy Clustering Validity Index Postprint

Authors: Geng Jiayi, Qian Xuezhong, Zhou Shibing

Date: 2018-05-02T00:00:00+00:00

Abstract

Fuzzy clustering is an important research topic in fields such as pattern recognition, machine learning, and image processing. The fuzzy C-means clustering algorithm is the most commonly used implementation of fuzzy clustering, which requires the number of clusters to be specified in advance before clustering a dataset. A new clustering validity index is proposed to validate the effectiveness of clustering results. This index defines three important feature measurements —compactness, separation, and overlap—from the perspectives of partition entropy, membership degree, and geometric structure. Based on this, a method for determining the optimal number of clusters is proposed. The new clustering validity index and traditional validity indices are experimentally validated on six artificial datasets and three real-world datasets. Experimental results demonstrate that the proposed index and method can effectively evaluate clustering results and are suitable for determining the optimal number of clusters for samples.

Full Text

Preamble

Journal Information: ChinaXiv Cooperative Journal, *Application Research of Computers*
Article: New Fuzzy Clustering Validity Index
Authors: Geng Jiayi, Qian Xuezhong, Zhou Shibing (School of Internet of Things Engineering, Jiangnan University, Wuxi, Jiangsu 214122)

Abstract

Fuzzy clustering is an important research area in pattern recognition, machine learning, and image processing. The fuzzy C-means clustering algorithm is the most commonly used implementation of fuzzy clustering, but it requires the number of clusters to be specified in advance. This paper proposes a new clustering validity index to evaluate clustering results. The index defines three

important feature measures—compactness, separation, and overlap degree—from the perspectives of partition entropy, membership degree, and geometric structure. Based on this, a method for determining the optimal number of clusters is proposed. The new clustering validity index and traditional indices are validated on six artificial datasets and three real datasets. Experimental results demonstrate that the proposed index and method can effectively evaluate clustering results and are suitable for determining the optimal number of clusters for samples.

Keywords: fuzzy C-means clustering; number of clusters; clustering validity index; fuzzy clustering

Funding: National Natural Science Foundation of China (61673193); Fundamental Research Funds for the Central Universities (JUSRP11235, JUSRP51635B)

Author Biographies: - Geng Jiayi (1992–), female, from Jinzhong, Shanxi, master' s student, research interests include pattern recognition and machine learning. - Qian Xuezhong (1967–), male, from Wuxi, Jiangsu, associate professor, master' s supervisor, research interests include data mining and machine learning. - Zhou Shibing (1972–), male, from Yancheng, Jiangsu, lecturer, Ph.D., research interests include pattern recognition and artificial intelligence.

---

0 Introduction

Clustering is the process of grouping samples without prior knowledge according to specific rules, where similar samples are grouped into the same class and dissimilar samples are assigned to different classes {[}{1}{]}. Clustering is divided into two main directions: traditional clustering and fuzzy clustering. Traditional clustering employs hard partitioning, where each sample must be clearly assigned to different subclasses—there are only two possibilities: belonging or not belonging. However, most real-world data exhibit uncertainty, and a sample may belong to multiple classes to varying degrees {[}{2}{]}. Therefore, Ruspini {[}{3}{]} introduced the concept of fuzzy partitioning, leading to the emergence of fuzzy clustering. Correspondingly, the range of membership degrees was extended from binary logic {0,1} to [0,1]. Compared with traditional clustering, fuzzy clustering better reflects the real world. The most commonly used algorithm for implementing fuzzy clustering is the fuzzy C-means algorithm (FCM) proposed by Dunn {[}{4}{]}. This algorithm minimizes the objective function through iteration. The FCM algorithm is simple in design and has a wide range of applications. However, the FCM algorithm requires cluster validity verification to determine the optimal number of clusters and evaluate the quality of classification results {[}{5}{]}.

Selecting an appropriate clustering validity index is a crucial step in cluster validity research {[}{6}{]}. Numerous clustering validity indices currently exist,

but due to the diverse structures of datasets, no single index is suitable for all types of datasets, and no index consistently outperforms others {[]{7}{]}. For instance, Piao Shangzhe et al. {[]{8}{]} listed several commonly used clustering validity indices in recent years, including those based on membership degree, intra-cluster compactness and inter-cluster separation, and those based on entropy and data structure. These indices only demonstrate their advantages on specific datasets and cannot be applied to all datasets. This paper addresses the limitations of existing fuzzy clustering validity indices by proposing a new clustering validity index. This index combines partition entropy, membership degree, and data structure to define compactness, separation, and overlap degree, which can overcome the effects of noise and overlap to accurately identify the optimal number of clusters. Experimental results demonstrate that the new index achieves good performance on both artificial and real datasets.

1.1 FCM Algorithm

The FCM algorithm is based on objective function optimization for fuzzy C-partitioning, which obtains uniform c fuzzy sets by optimizing the objective function {[]{9}{]}. The objective function consists of a combination of membership degrees and deviations of samples from cluster centers. Through iterative minimization of the objective function, the process terminates when the number of iterations exceeds a specified value or the difference in the objective function falls below a threshold. FCM requires prior initialization of cluster prototypes and specification of the number of clusters.

Assume a dataset has n samples, each being p-dimensional. The objective function is:

$$J(U,V,X)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}\|x_k-v_i\|^2$$

Subject to the constraints on the membership matrix:

$$0\le u_{ij}\le 1,\quad 1\le i\le c,\quad 1\le j\le n$$

$$\sum_{i=1}^{c}u_{ij}=1,\quad 1\le j\le n$$

$$0<\sum_{j=1}^{n}u_{ij}<n,\quad 1\le i\le c$$

where: c represents the number of clusters; m represents the fuzziness parameter in the range [1, ∞]; U is a c×N matrix representing the membership degree

of samples belonging to fuzzy subsets; V is a c×p matrix representing cluster prototypes. The FCM algorithm minimizes the objective function by iteratively updating the cluster prototypes V and membership matrix U. The update formulas are:

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m}, \quad 1 \le i \le c$$

$$u_{ik} = \frac{\|x_k - v_i\|^{-2/(m-1)}}{\sum_{j=1}^{c} \|x_k - v_j\|^{-2/(m-1)}}, \quad 1 \le i \le c, \quad 1 \le k \le n$$

The steps of the FCM algorithm are as follows: a) Specify parameters: number of clusters C, fuzziness parameter m, maximum number of iterations, and iteration termination condition. b) Initialize cluster prototypes and update the fuzzy membership matrix U. c) Update the fuzzy cluster prototype matrix V. d) If the maximum number of iterations is reached or the difference in the objective function is below the threshold, stop; otherwise, return to step b.

1.2 Traditional Clustering Validity Indices

1) **PC {[}{10}{]}**: The partition coefficient PC has a simple form and is easy to compute, but it only considers the compactness of each cluster and lacks direct connection to the geometric structure of the data. It exhibits a monotonic trend as the number of clusters changes. These deficiencies directly prevent the index from validating clustering effectiveness for partitions with numerous small clusters and complex datasets.

2) **MPC {[}{11}{]}**: This index optimizes the monotonic decreasing trend problem of the partition coefficient PC, but it does not improve other aspects of PC' s deficiencies. The index performs poorly on artificial datasets.

3) **PE {[}{12}{]}**: The partition entropy PE index is simple and has low computational complexity, but it suffers from several problems: it only considers the compactness of each cluster, lacks connection to the geometric structure of the dataset, and exhibits a monotonic trend. The index performs well only on well-separated datasets and poorly on noisy and overlapping datasets.

4) **XB {[}{13}{]}**: This index considers both membership degrees and geometric structure of the data, where compactness is the sum of distances from all samples to cluster centers, and separation is the minimum distance between cluster centers. The index has two drawbacks: when c→n, the XB index becomes 0; when m→∞, XB→∞. In both cases, the index loses stability and cannot determine the optimal number of clusters.

5) **UV {[}{14}{]}**: This index introduces an exponential function to measure the distance between data and centers, which can overcome the influence of noise to some extent compared with Euclidean distance. However, since the index only considers cluster compactness and separation without accounting for the significant impact of overlap degree on classification, its performance is not ideal on overlapping datasets.

6) **FM {[}{15}{]}**: This index considers both partition entropy and fuzzy partition factor to define cluster compactness and separation. Since it uses the distance between the two nearest cluster centers as separation, the index performs poorly on noisy data.

---

2 New Clustering Validity Index

The quality of a clustering validity index directly affects the quality of the final clustering results. The new clustering validity index is composed of three components: compactness, overlap degree, and separation degree. Compactness is represented by intra-cluster distance, separation by minimum membership degree, and overlap degree by a combination of membership degree and partition entropy. Good clustering corresponds to smaller compactness, smaller overlap degree, and larger separation degree. This index fully considers the overall information of the dataset and can accurately determine the optimal number of clusters.

2.1 Compactness

Compactness represents the degree of concentration of samples within a class. To explain the relevant concepts, schematic diagrams are used for illustration. [Figure 1: see original paper] shows the distance from all samples of class i to the cluster center of that class; smaller values indicate that intra-class samples are closer to the cluster center, demonstrating the structural relationship between intra-class samples and the cluster center. [Figure 2: see original paper] shows the pairwise distances between all sample data in class i; smaller values indicate that data within the class are tighter, demonstrating the overall structural information of intra-class sample data. Intra-class compactness combines these two aspects to leverage their respective advantages. Clearly, the minimum value of Var(c,U) indicates that data points within the class are close to each other, resulting in higher compactness.

Definition 1: Define vs(c,U) as the average distance from all samples of class i to the center of class i:

$$vs(c,U) = \frac{\sum_{k=1}^{n(i)} \|x_k^{(i)} - v_i\|}{n(i)}$$

where $x_k^{(i)}$ represents the k-th sample of class i, and $n(i)$ represents the number of samples in the i-th cluster.

Definition 2: Define vd(c,U) as the average pairwise distance between all samples in class i:

$$vd(c,U) = \frac{\sum_{k=1}^{n(i)} \sum_{h=1}^{n(i)} \|x_k^{(i)} - x_h^{(i)}\|^2}{(n(i) - n(i))/2}$$

where the denominator represents the total number of pairwise distances among all samples in class i.

Definition 3: Define intra-class compactness Var(c,U) as the sum of the previous two components (average distance from samples to center and average distance between intra-class samples):

$$Var(c,U) = vs(c,U) \times vd(c,U)$$

2.2 Separation Degree

Definition 4: Define S_{ij}^k as the minimum membership degree of the k-th sample belonging to classes i and j:

$$S_{ij}^k = \min(u_{ik}, u_{jk}), \quad k = 1, 2, ..., n, \quad i \neq j$$

Definition 5: Define overall separation degree Sep(c,U) as the negative sum of S_{ij}^k:

$$Sep(c,U) = 1 - \frac{\sum_{i=1}^{c-1} \sum_{j=i+1}^{c} \sum_{k=1}^{n} S_{ij}^k}{c \times n}$$

Separation degree represents the degree of separation between two fuzzy clusters. Most indices measure separation by calculating distances between cluster centers, which cannot reflect the overall shape of sample distribution and may produce biases for noisy data. For example, in [Figure 3: see original paper], two classes have the same distance but may have different separability. Classes AB and AC have equal center distances, but AB is clearly more separated than AC. The new index draws on the separation measure proposed by Chen et al. {[}{16}{]}, using the minimum fuzzy membership value of a sample data point relative to two classes as the separation measure. The closer the k-th sample is to the center of one class, the closer its membership degree is to 1 for that class and to 0 for the other class, making the value from Definition 4 approach 0. In this case, inter-class fuzziness is smaller and the classes are more separated. The overall separation degree sums the values from Definition 4 and takes the negative, where larger values indicate lower fuzziness of data points relative to classes, enabling clearer partitioning into clusters and better separation.

2.3 Overlap Degree

Definition 6: Define C_{ij}^k as the product of membership degrees of the k-th sample belonging to classes i and j:

$$C_{ij}^k = (u_{ik} \times u_{jk})^2, \quad k = 1, 2, ..., n, \quad i \neq j$$

Definition 7: Define overall overlap degree Cop(c,U) as the sum of C_{ij}^k combined with entropy:

$$Cop(c, U) = \frac{\sum_{i=1}^{c-1} \sum_{j=i+1}^{c} \sum_{k=1}^{n} C_{ij}^k \times f(x)}{n}$$

where

$$f(x) = -\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik} \log u_{ik}$$

Overlap degree measures the degree of overlap between two classes with unclear boundaries. The overlap between two classes is defined as the product of squared membership degrees. When the division between two categories is clear and membership degrees differ significantly, the product value is smaller, resulting in clearer class partitioning and clustering results. The overlap degree reaches its maximum when the membership degree of the k-th sample relative to each class is 1/c. Combining this with entropy better reflects the fuzziness and uncertainty of the partitioning result; smaller values indicate lower uncertainty and less required information, making the classification more reliable. This serves as a weighting evaluation metric. Clearly, smaller overall overlap degree values indicate clearer partitioning between two classes with less overlap.

2.4 Normalization

Since compactness, separation degree, and overlap degree have different dimensions, normalization is required. The index values corresponding to each number of clusters are divided by the maximum index value, transforming each measure' s range to [0,1]. The results can be expressed as:

$$Var_n(c, U) = \frac{Var(c, U)}{Var_{\max}}, \quad Var_{\max} = \max(Var(c, U))$$

$$Sep_n(c, U) = \frac{Sep(c, U)}{Sep_{\max}}, \quad Sep_{\max} = \max(Sep(c, U))$$

$$Cop_n(c, U) = \frac{Cop(c, U)}{Cop_{\max}}, \quad Cop_{\max} = \max(Cop(c, U))$$

2.5 Clustering Validity Index

By combining partition entropy, membership degree, and geometric structure —important features in fuzzy clustering—a new clustering validity index is constructed. From the perspective of compactness, we want Var(c,U) to be as small as possible, indicating tighter intra-class distances. From the perspective of separation, we want Sep(c,U) to be as large as possible, indicating more dispersed inter-class distances. From the perspective of overlap, we want Cop(c,U) to be as small as possible, indicating minimal overlap. Clearly, smaller W(c,U) values indicate that data points are clearly assigned to clusters, representing the best clustering performance. The optimal number of clusters should match the true structure of the dataset, and finding this number is the primary task of a clustering validity index. This index can accurately determine the optimal number of clusters on both noisy and overlapping datasets.

The new index is defined as:

$$W(c,U) = Var_n(c,U) + \frac{Cop_n(c,U)}{Sep_n(c,U)}$$

---

3 Algorithm for Determining the Optimal Number of Clusters

This paper proposes a new algorithm for determining the optimal number of clusters based on the FCM algorithm and the W clustering validity index, which solves the problem of FCM requiring prior specification of the number of clusters. The steps are as follows:

a) Initialize the search range for the number of clusters as [C_min, C_max].
b) Increment c by 1, call the FCM algorithm, and use the optimal solution (U,V) obtained from FCM to compute the W index.
c) Calculate and store the clustering validity index values.
d) If c < C_max, set c = c + 1 and return to step 2; otherwise, proceed to step 5.
e) Select the c corresponding to the minimum index value as the optimal number of clusters.
f) Output the optimal number of clusters and the index values.

---

4 Experimental Results

To verify the effectiveness of the new clustering validity index, the proposed index (W) and existing indices (PC, MPC, PE, XB, UV, FM) were applied to six artificial datasets and three real datasets to observe their clustering performance. The search range for the number of clusters was [2, C_max], where C_max =

√n. Euclidean distance was used for all distance metrics in the indices. While theoretical guidance is lacking for the optimal value of parameter m, Pal and Bezdek {[}{17}{]} suggested that FCM clustering produces the best results when m is in [1.5, 2.5]. Experiments were first conducted with m = 2, and the robustness of the new index was observed under different fuzzy weighting values of m.

4.1.1 Artificial Datasets

[Figure 4: see original paper] shows the distribution structures of the artificial datasets. DS1, DS2, and DS3 are Gaussian distribution datasets, while DS4, DS5, and DS6 are uniform distribution datasets. DS1 and DS4 have clear separation between classes. DS2 and DS5 contain 150 noise-contaminated data points. In DS3, three class samples have significant overlap with each other, while the other two classes are well-separated. In DS6, all five class samples have certain overlaps with each other.

presents the specific numerical information of the artificial datasets and the optimal number of clusters determined by each validity index. For clearer illustration, [Figure 5: see original paper] details the cluster number-index relationship plots for DS2 and DS3 datasets. For well-separated datasets DS1 and DS4, all indices work effectively. For noisy datasets: on DS2, PE and FM identified 2 clusters as optimal (with 4 as the second best), XB incorrectly identified 6 clusters, while the remaining indices correctly determined 4 clusters; on DS5, only MPC and W correctly identified 3 clusters. For overlapping datasets: DS3 and DS6, due to extensive overlapping regions, traditional clustering validity indices lose their discriminative ability, with only the W index correctly identifying the optimal number of clusters as 5 for both datasets.

4.1.2 Real Datasets

The real datasets were obtained from the public UCI database, a standard test database for machine learning proposed by the University of California, Irvine.

a) **Iris dataset**: Contains 3 classes (Iris Setosa, Iris Versicolour, Iris Virginica). In this dataset, two classes are almost indistinguishable while the third cluster is well-separated. Therefore, the optimal number of clusters should be 3, with 2 as the second-best option—both consistent with the dataset structure. Traditional indices identified 2 as optimal, while the new index correctly identified 3.

b) **Wdbc dataset**: Contains 2 classes (Malignant, Benign). Features were computed from digitized images of fine needle aspirate (FNA) of breast masses, describing characteristics of cell nuclei present in the images. Nuclear feature extraction is used for breast tumor diagnosis. All indices correctly identified the optimal number of clusters as 2.

c) **Seeds dataset**: Contains 3 classes representing three different varieties

of wheat kernels (Kama, Rosa, and Canadian). Only the new clustering validity index correctly identified 3 clusters, while all other indices incorrectly identified 2 as optimal.

provides specific numerical information for the real datasets and the optimal number of clusters determined by each validity index. For clearer illustration, [Figure 6: see original paper] details the cluster number-index relationship plots for the Iris and Seeds datasets. Experimental results show that only the new clustering validity index can correctly determine the optimal number of clusters across all artificial and real datasets, demonstrating good performance on both noisy and overlapping data.

4.2 Comparison of m Values

The parameter m is the fuzzy weighting exponent that controls the fuzziness of clustering results. Its introduction extends traditional clustering to fuzzy clustering. Based on existing experience, m is generally selected from the range [1.5, 2.5]. The clustering validity indices were applied to the aforementioned datasets under five different m values: 1.5, 1.7, 2, 2.3, and 2.5. Experimental results show that PC, MPC, PE, XB, and UV produce varying clustering results as m changes, while only FM and W remain unchanged with m, demonstrating robustness to m. In terms of clustering accuracy, only the newly proposed W index achieves 100% correctness.

Experiments demonstrate that the new index can produce good results under different m values, showing strong reliability and robustness. PC, PE, and MPC indices lack direct connection to data structure, causing the number of clusters corresponding to the optimal validity index to deviate from actual conditions. The XB index loses reliability when m and c increase to certain extents. UV and FM only consider compactness and separation without accounting for overlap degree. The new index compensates for these shortcomings of traditional indices and demonstrates strong adaptability.

through show the optimal number of clusters for W, PC, MPC, PE, XB, UV, and FM indices under different fuzzy weighting exponents.

---

5 Conclusion

To address the deficiencies of existing indices, this paper proposes the W clustering validity index. Experimental results on artificial and real datasets demonstrate that the W index can make correct judgments in the presence of noise and inter-class overlap, with minimal correlation to the fuzzy weighting exponent and stable performance. Since the FCM algorithm primarily processes cluster-shaped distribution data, its effectiveness is poor for non-cluster-shaped distribution data. Future research will focus on validity issues for non-cluster-shaped distribution data.

References

[1] Zalik K R. Cluster validity index for estimation of fuzzy clusters of different [J]. Pattern Recognition, 2010, 43 (10): 3374-3390.
[2] Yang Lei. Extending information-theoretic validity indices for fuzzy clustering [J]. IEEE Trans on Fuzzy Systems, 2017, 25 (4): 1013-1018.
[3] Ruspini E H. A New Approach to Clustering [J]. InfCont, 1969, 15 (1): 22-32.
[4] Dunn J C. A Fuzzy relative of the ISODATA process its use in detecting compact well-separated clusters[J]. Journal of Cybernetics, 1974, 3(3): 32-57.
[5] Le Capitaine H, Carl Freli. A cluster-validity index combining an overlap measure and a separation measure based on fuzzy-aggregation operators [J]. IEEE Trans on Fuzzy Systems, 2011, 19 (3): 580-588.
[6] Bezdek J C, Life Fellow. The generalized c index for internal fuzzy cluster validity [J]. IEEE Trans on Fuzzy Systems. 2016, 24 (6): 1500-1512.
[7] 高新波, 谢维信. 模糊聚类理论发展及应用的研究进展 [J]. 科学通报, 1999, 44 (21): 2241-2251.
[8] 朴尚哲, 超木日力格, 于剑. 模糊 C 均值算法的聚类有效性评价 [J]. 模式识别与人工智能, 2015, 28 (5): 452-461.
[9] Tas demir K. A validity index for prototype-based clustering of data sets with complex cluster structures [J]. IEEE Trans on Systems, 2011, 41 (4): 1039-1053.
[10] 范九伦. 基于模糊熵的聚类有效性函数 _ 范九伦 [J]. 模式识别与人工智能, 2001, 14 (4): 390-394.
[11] Liliane Silva. An Interval-based framework for fuzzy clustering applications [J]. IEEE Trans on Systems, 2015, 23 (6): 2174-2187.
[12] 毕凯. 基于模糊测度和证据理论的模糊聚类集成方法 [J]. 控制与决策, 2015, 30 (5): 823-830.
[13] Xie X L, Beni G. A validity measure for fuzzy clustering [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1991, 13 (8): 841-847.
[14] 张宇献, 刘通. 基于改进划分系数的模糊聚类有效性函数 [J]. 沈阳工业大学学报, 2014, 36 (4): 431-435.
[15] 孟令奎, 胡春春. 基于模糊划分测度的聚类有效性指标 [J]. 计算机工程, 2007, 33 (11): 15-17.
[16] Chen M Y, Linkens D A. Rule-base self-generation and simplification for data-driven fuzzy models [J]. Fuzzy Sets and Systems, 2004, 142 (2): 243-265.
[17] Pal N R, Bezdek J C. On Cluster Validity for the fuzzy C-Means model [J]. IEEE Trans on Fuzzy Systems, 1995, 3 (3): 370-379.

Note: Figure translations are in progress. See original paper for figures.