A New Validity Measure for Heuristic Possibilistic Clustering

Dmitri A. Viattchenin

United Institute of Informatics Problems of the National Academy of Sciences of Belarus
Surganov Str. 6, 220012 Minsk, Belarus
viattchenin@mail.ru

Frank Klawonn and Katharina Tschumitschew

Ostfalia University of Applied Sciences
Salzdahlumer Str. 46/48, D-38302 Wolfenbuettel, Germany
f.klawonn@ostfalia.de, k.tschumitschew@ostfalia.de

Abstract

A heuristic approach to possibilistic clustering is the effective tool for the data analysis. The approach is based on the concept of allotment among fuzzy clusters. To establish the number of clusters in a data set, a validity measure is proposed in this paper. An illustrative example of application of the proposed validity measure to the Anderson's Iris data is given. A comparison of the validity measure with some well-known cluster validity indices for objective function-based fuzzy clustering algorithms is given. Preliminary conclusions are formulated.

1 Introduction

The objective of a fuzzy clustering algorithm is to partition a data set $X = \{x_1, ..., x_n\}$ into c homogeneous fuzzy clusters. The most widely used fuzzy clustering algorithm is the FCM-algorithm which was proposed by Bezdek (1981). The FCM-algorithm is the basis of the family of fuzzy clustering algorithms. The family of objective function-based fuzzy clustering algorithms includes

- fuzzy c-lines algorithm (FCL);
- fuzzy c-rings algorithm (FCR);
- fuzzy c-shells algorithm (FCS);
- fuzzy c-rectangular shells algorithm (FCRS).

These and other well-known fuzzy clustering algorithms were proposed by different authors and they are considered by Höppner, Klawonn, Kruse and Runkler (1999) in detail. Moreover, some other fuzzy clustering algorithms were also proposed.

All fuzzy clustering algorithms require the user to pre-define the number of clusters, c. However, it is not always possible to know the number of clusters in advance. Different fuzzy partitions are obtained at different values of c. Thus, a methodology of evaluation of fuzzy partitions is required to validate each fuzzy partition to obtain the optimal number of clusters, c.

Many cluster validity criterion have been proposed for validating fuzzy partition. In particular, the partition coefficient (V_{pc}), the partition entropy (V_{pe}), the Xie-Beni index (V_{XB}), the Fukuyama-Sugeno index (V_{FS}) are well-known validity measures, and these criteria are considered by Höppner, Klawonn, Kruse and Runkler (1999).

However, the condition of fuzzy partition is very difficult from essential positions. That is why a possibilistic approach to clustering was proposed by Krishnapuram and Keller (1993) and developed by other researchers. This approach can be considered as a way in the optimization approach in fuzzy clustering because all methods of possibilistic clustering are objective function-based methods. A concept of possibilistic partition is a basis of possibilistic clustering methods and membership values can be interpreted as the values of typicality degree.

An outline for a new heuristic method of fuzzy clustering was presented by Viattchenin (2004), where concepts of fuzzy α-cluster and allotment among fuzzy α-clusters were introduced and a basic version of direct fuzzy clustering algorithm was described. The basic version of direct fuzzy clustering algorithm requires that the number c of fuzzy α-clusters be fixed. That is why the basic version of the algorithm, which was proposed by Viattchenin (2004), can be called the D-AFC(c)-algorithm. Moreover, the allotment of elements of the set of classified objects among fuzzy α-clusters can be considered as a special case of possibilistic partition. So, the D-AFC(c)-algorithm can be considered as a direct algorithm of possibilistic clustering. The fact was demonstrated by Viattchenin (2007).

The results of application of the D-AFC(c)-algorithm to the Anderson's (1935) Iris data are considered by Viattchenin (2006) and the results shows that the D-AFC(c)-algorithm is a precise and effective numerical procedure for solving classification problems. However, a unique validity measure for the D-AFC(c)-algorithm was proposed by Viattchenin, Damaratski, and Novikau (2009) and the proposed linear measure of fuzziness of the allotment is not effective in every classification problem. So, the main goal of this paper is a proposition of a new validity measure for the D-AFC(c)-algorithm. The contents of this paper is as follows: in the second section basic concepts of the clustering method are considered, the linear measure of fuzziness of the allotment is presented and a new cluster validity index for the D-AFC(c)-algorithm is proposed, in the third section an example of application of the D-AFC(c)-algorithm with the proposed validity measure to the Anderson's (1935) Iris data set is given in comparison with the linear measure of fuzziness of the allotment and some validity criteria for validating fuzzy partition, in the fourth section methods of the data preprocessing are considered and numerical in comparison with the compactness and separation index for objective function-based fuzzy clustering algorithms, in the fifth section some final remarks are stated.

2

2 A heuristic method of possibilistic clustering

The basic concepts of the heuristic method of possibilistic clustering are considered in the first subsection. The linear measure of fuzziness of the allotment is presented in the second subsection and a new validity measure is proposed in the third subsection of the section.

2.1 Basic concepts

Let us remind the basic concepts of the D-AFC(c)-algorithm. The concept of fuzzy tolerance is the basis for the concept of fuzzy α-cluster. That is why definition of fuzzy tolerance must be considered in the first place.

Let $X = \{x_1,...,x_n\}$ be the initial set of elements and $T : X \times X \to [0,1]$ some binary fuzzy relation on X with $\mu_T(x_i, x_j) \in [0,1]$, $\forall x_i, x_j \in X$ being its membership function. Fuzzy tolerance is the fuzzy binary intransitive relation which possesses the symmetricity property

$$\mu_T(x_i, x_j) = \mu_T(x_j, x_i), \ \forall x_i, x_j \in X, \qquad (1)$$

and the reflexivity property

$$\mu_T(x_i, x_i) = 1, \ \forall x_i \in X. \qquad (2)$$

Let $X = \{x_1,...,x_n\}$ be the initial set of objects. Let T be a fuzzy tolerance on X and α be α-level value of T, $\alpha \in (0,1]$. Columns or lines of the fuzzy tolerance matrix are fuzzy sets $\{A^1,...,A^n\}$. Let $\{A^1,...,A^n\}$ be fuzzy sets on X, which are generated by a fuzzy tolerance T. The α-level fuzzy set $A^l_{(\alpha)} = \{(x_i, \mu_{A^l}(x_i)) \mid \mu_{A^l}(x_i) \geq \alpha, l \in [1,n]\}$ is fuzzy α-cluster or, simply, fuzzy cluster. So $A^l_{(\alpha)} \subseteq A^l$, $\alpha \in (0,1]$, $A^l \in \{A^1,\ldots,A^n\}$ and μ_{li} is the membership degree of the element $x_i \in X$ for some fuzzy cluster $A^l_{(\alpha)}$, $\alpha \in (0,1]$, $l \in [1,n]$. Value of α is the tolerance threshold of fuzzy clusters elements.

The membership degree of the element $x_i \in X$ for some fuzzy cluster $A^l_{(\alpha)}$, $\alpha \in (0,1]$, $l \in [1,n]$ can be defined as a

$$\mu_{li} = \begin{cases} \mu_{A^l}(x_i), & x_i \in A^l_\alpha \\ 0, & otherwise \end{cases}, \qquad (3)$$

where an α-level $A^l_\alpha = \{x_i \in X \mid \mu_{A^l}(x_i) \geq \alpha\}$, $\alpha \in (0,1]$ of a fuzzy set A^l is the support of the fuzzy cluster $A^l_{(\alpha)}$. So, condition $A^l_\alpha = Supp(A^l_{(\alpha)})$ is met for each fuzzy cluster $A^l_{(\alpha)}$, $\alpha \in (0,1]$, $l \in [1,n]$. Membership degree can be interpreted as a degree of typicality of an element to a fuzzy cluster. The value of a membership function of each element of the fuzzy cluster in the sense of (3) is the degree of similarity of the object to some typical object of fuzzy cluster.

Let T is a fuzzy tolerance on X, where X is the set of elements, and $\{A^1_{(\alpha)},...,A^n_{(\alpha)}\}$ is the family of fuzzy clusters for some $\alpha \in (0,1]$. The point $\tau^l_e \in A^l_\alpha$, for which

$$\tau^l_e = \arg\max_{x_i} \mu_{li}, \ \forall x_i \in A^l_\alpha \qquad (4)$$

is called a typical point of the fuzzy cluster $A^l_{(\alpha)}$, $\alpha \in (0,1]$, $l \in [1,n]$. A fuzzy cluster can have several typical points. So, symbol e is the index of the typical point.

Let $R^\alpha_z(X) = \{A^l_{(\alpha)} \mid l = \overline{1,c}, 2 \leq c \leq n, \alpha \in (0,1]\}$ be a family of fuzzy clusters for some value of tolerance threshold α, $\alpha \in (0,1]$, which are generated by some fuzzy tolerance T on the initial set of elements $X = \{x_1,...,x_n\}$. If condition

$$\sum_{l=1}^{c}\mu_{li} > 0,\ \forall x_i \in X \qquad (5)$$

is met for all fuzzy clusters $A_{(\alpha)}^l \in R_z^\alpha(X)$, $l = \overline{1,c}$, $c \le n$, then the family is the allotment of elements of the set $X = \{x_1,...,x_n\}$ among fuzzy clusters $\{A_{(\alpha)}^l, l = \overline{1,c}, 2 \le c \le n\}$ for some value of the tolerance threshold α.

It should be noted that several allotments $R_z^\alpha(X)$ can exist for some tolerance threshold α. That is why symbol z is the index of an allotment.

The condition (5) requires that every object x_i, $i = 1,\ldots,n$ must be assigned to at least one fuzzy cluster $A_{(\alpha)}^l$, $l = \overline{1,c}$, $c \le n$ with the membership degree higher than zero. The condition $2 \le c \le n$ requires that the number of fuzzy clusters in each allotment $R_z^\alpha(X)$ must be more than two. Obviously, the definition of the allotment among fuzzy clusters (5) is similar to the definition of the possibilistic partition. So, the allotment among fuzzy clusters can be considered as the possibilistic partition and fuzzy clusters in the sense of (3) are elements of the possibilistic partition.

If condition

$$\sum_{l=1}^{c} card(A_\alpha^l) \ge card(X),\ \forall A_{(\alpha)}^l \in R_z^\alpha(X), \qquad (6)$$

and condition

$$card(A_\alpha^l \cap A_\alpha^m) \le w,\ \forall A_{(\alpha)}^l, A_{(\alpha)}^m,\ l \ne m, \qquad (7)$$

are met for all fuzzy clusters $A_{(\alpha)}^l, l = \overline{1,c}$ of some allotment $R_z^\alpha(X) = \{A_{(\alpha)}^l \mid l = \overline{1,c}, c \le n\}$ then the allotment is the allotment among particularly separate fuzzy clusters and $0 \le w \le n$ is the maximum number of elements in the intersection area of different fuzzy clusters.

Obviously, if $w = 0$ in conditions (6) and (7) then the intersection area of any pair of different fuzzy cluster is an empty set and fuzzy clusters are fully separate fuzzy clusters.

Detection of fixed c number of fuzzy clusters can be considered as the aim of classification. So, the allotment $R_z^\alpha(X) = \{A_{(\alpha)}^l \mid l = \overline{1,c}\}$ among the given number c of fuzzy clusters and the corresponding value of tolerance threshold α are the results of classification.

A plan of the D-AFC(c)-algorithm is presented, for example, by Viattchenin (2007).

2.2 The linear measure of fuzziness of the allotment

The linear measure of fuzziness of the allotment which is the validity measure for the D-AFC(c)-algorithm was defined by Viattchenin, Damaratski, and Novikau (2009) as follows:

$$V_{LMF}\left(R^*(X);c\right) = \sum_{A_{(\alpha)}^l \in R^*(X)} I_L(A_{(\alpha)}^l), \qquad (8)$$

where $I_L(A_{(\alpha)}^l)$ is a modification of the linear index of fuzziness which was defined by Viattchenin (2006) as

$$I_L(A_{(\alpha)}^l) = \frac{2}{n_l} \cdot d_H(A_{(\alpha)}^l, \underline{A}_{(\alpha)}^l), \qquad (9)$$

where $n_l = card(A_\alpha^l)$, $A_{(\alpha)}^l \in R^*(X)$ is the number of objects in the fuzzy cluster $A_{(\alpha)}^l$ and $d_H(A_{(\alpha)}^l, \underline{A}_{(\alpha)}^l)$ is the Hamming distance

$$d_H(A_{(\alpha)}^l, \underline{A}_{(\alpha)}^l) = \sum_{x_i \in A_\alpha^l} \left|\mu_{li} - \mu_{\underline{A}_{(\alpha)}^l}(x_i)\right| \qquad (10)$$

between the fuzzy cluster $A_{(\alpha)}^l$ and the crisp set $\underline{A}_{(\alpha)}^l$ nearest to the fuzzy cluster $A_{(\alpha)}^l$. The membership function of the crisp set $\underline{A}_{(\alpha)}^l$ can be defined as

$$\mu_{\underline{A}_{(\alpha)}^l}(x_i) = \begin{cases} 0, & \mu_{A_{(\alpha)}^l}(x_i) \le 0.5, \\ 1, & \mu_{A_{(\alpha)}^l}(x_i) > 0.5, \end{cases} \quad \forall x_i \in A_\alpha^l, \qquad (11)$$

where $\alpha \in (0,1]$.

The fuzziness of the allotment $R^*(X)$ depends on the size of each fuzzy cluster. Using $V_{LMF}(R^*(X);c)$, the optimal number c of fuzzy clusters can be obtained by maximizing the index (8) value.

2.3 A new cluster validity measure

From other hand, a density of fuzzy clusters can be taken into account for the validating allotment. The density of fuzzy cluster was defined by Viattchenin (2006) as follows:

$$D(A^l_{(\alpha)}) = \frac{1}{n_l} \sum_{x_i \in A^l_\alpha} \mu_{li}\ , \tag{12}$$

where $n_l = card(A^l_\alpha)$, $A^l_{(\alpha)} \in R^*(X)$ and membership degree μ_{li} is defined by formula (3). It is obvious that condition

$$0 < D(A^l_{(\alpha)}) \leq 1, \tag{13}$$

is met for each fuzzy cluster $A^l_{(\alpha)}$ in $R^*(X)$. Moreover, $D(A^l_{(\alpha)}) = 1$ for a crisp set $A^l_{(\alpha)} \in R^*(X)$ for any tolerance threshold α, $\alpha \in (0,1]$. The density of fuzzy cluster shows an average membership degree of elements of a fuzzy cluster.

The density of fuzzy cluster (12) can be considered as the basis for a validity measure, too. The validity measure must be taking into account the compactness of fuzzy clusters which is characterized by their density. The density of each fuzzy cluster $A^l_{(\alpha)} \in R^*(X)$ is increasing with increasing of the number c of fuzzy clusters. So, for $c \to n$ we have $D(A^l_{(\alpha)}) \to 1$ for all $A^l_{(\alpha)}$, $l \in \{1, \ldots, c\}$. Moreover, for $c \to n$ we have $\alpha \to 1$. Thus, the value of the tolerance threshold α must be taken into account. So, the validity measure can be defined as the ratio of the sum of densities of fuzzy clusters of some allotment to the number of fuzzy clusters minus the value of the tolerance threshold α. However, a case of particularly separate fuzzy clusters must be taken into account. That is why a total number of objects, $\breve{n}$, in intersection areas of each pair of fuzzy clusters must be calculated.

Thus, the measure of compactness of the allotment can be defined in the following way:

$$V_{MC}(R^*(X);c) = \begin{cases} \dfrac{\sum\limits_{A^l_{(\alpha)} \in R^*(X)} D(A^l_{(\alpha)})}{c} - \alpha, & \textit{for fully separate fuzzy clusters} \\ \breve{n} \cdot \left(\dfrac{\sum\limits_{A^l_{(\alpha)} \in R^*(X)} D(A^l_{(\alpha)})}{c} - \alpha \right), & \textit{for particularly separate fuzzy clusters} \end{cases}, \tag{14}$$

where $\breve{n}$ is the total number of elements in all intersection areas of different fuzzy clusters. Note that the value of $V_{MC}(R^*(X);c)$ for fully separate fuzzy clusters will be equal to the value for particularly separate fuzzy clusters in the case of $\breve{n} = 1$. The measure of compactness of the allotment $V_{MC}(R^*(X);c)$ increases when c is closer to n. Solving $\min\limits_c \left(V_{MC}\left(R^*(X);c\right)\right)$, $c = 2, \ldots, c_{\max}$, $c_{\max} \leq n-1$, is assumed to produce valid clustering of the initial data set X.

3 An illustrative example

The Anderson's (1935) Iris data set represents different categories of Iris plants having four attribute values. The four attribute values represent the sepal length, sepal width, petal length and petal width measured for 150 irises. It has three classes Setosa, Versicolor and Virginica, with 50 samples per class. The problem is to classify the plants into three subspecies on the basis of this information. The Anderson's Iris data can be presented as a matrix of attributes $\hat{X}_{n \times m} = [\hat{x}^t_i]$, $i = 1, \ldots, 150$, $t = 1, \ldots, 4$, where the value $\hat{x}^t_i$ is the value of the t-th attribute for i-th object. The method of the data preprocessing is described by Viattchenin (2006). The data were preprocessed using the squared normalized Euclidean distance (Kaufmann, 1975). So, the matrix of fuzzy tolerance $T = [\mu_T(x_i, x_j)]$, $i, j = 1, \ldots, n$ was obtained.

We applied the D-AFC(c)-algorithm to the matrix of fuzzy tolerance for $c = 2, \ldots, c_{\max} = 5$. So, we calculated the values of the linear measure of fuzziness of the allotment and the measure of compactness of the allotment for different c values and we plotted these validity measures in Figure 1 and Figure 2.

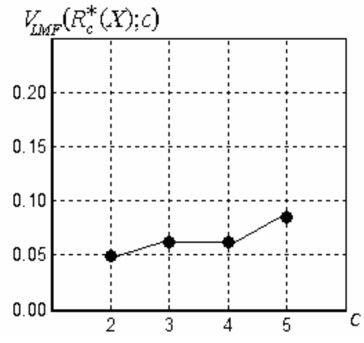

Figure 1: Plot of the linear measure of fuzziness of the allotment as a function of the number of clusters.

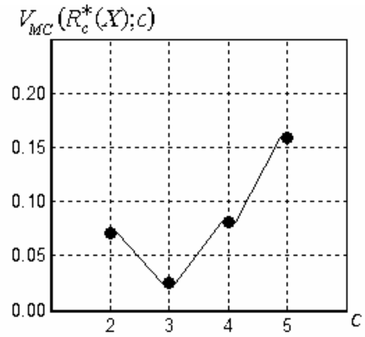

Figure 2: Plot of the measure of compactness of the allotment as a function of the number of clusters.

By executing the D-AFC(c)-algorithm for $c = 2, \ldots, c_{\max} = 5$, we obtain that the optimal cluster number c is chosen at $c = 5$ for the linear measure of fuzziness of the allotment. However, the number of fuzzy clusters $c = 3$ corresponds to the first maximum for the validity measure. From other hand, the measure of compactness of the allotment finds the optimal cluster number c at $c = 3$. Allotments among fully separated fuzzy clusters were obtained for $c = 2$ and $c = 3$. The value of the total number of elements in intersection areas $\breve{n}$ is equal 10 for the allotment among four particularly separate fuzzy clusters and for $c = 5$ we have $\breve{n} = 17$. So, the result of the proposed validity measure is seems as appropriate.

To demonstrate the effectiveness of the proposed validity measure, we conducted extensive comparison with some cluster validity indices for validating fuzzy partition. The partition coefficient (V_{pc}), the partition entropy (V_{pe}), the Xie-Beni index (V_{XB}), the Fukuyama-Sugeno index (V_{FS}) were selected for the comparison. The numbers of clusters yielded by all the validity indices for the Anderson's Iris data set are given in Table 1.

Table 1: Values of *c* preferred by each cluster validity index for the Iris data set.

The validity measure	***c***
V_{pc}	2
V_{pe}	2
V_{XB}	2
V_{FS}	3

Note that most validity measures reported in the literature provides two clusters for this data (Bouguessa, Wang, Sun, 2006).

4 Conclusions

A new cluster validity measure for the heuristic D-AFC(c)-algorithm of possibilistic clustering is introduced in the paper and a numerical experiment confirmed its utility. Thus, the result of application of the proposed validity measure to the data set shows that the validity measure is an effective tool for solving the classification problem.

Acknowledgements

The investigation is partially supported by the German Academic Exchange Service. Authors are grateful to Mr. Aliaksandr Damaratski for elaborating experimental software.

References

Anderson, E. (1935) "The Irises of the Gaspe Peninsula." Bulletin of the American Iris Society, Vol.59, pp. 2-5.

Bezdek, J.C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms.* New York: Plenum Press.

Bouguessa, M., S. Wang and H. Sun. (2006) "An Objective Approach to Cluster Validation." Pattern Recognition Letters, Vol.27, pp. 1419-1430.

Höppner, F., F. Klawonn, R. Kruse and T. Runkler. (1999) *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition.* Chichester: Wiley Intersciences.

Kaufmann, A. (1975) *Introduction to the Theory of Fuzzy Subsets.* New York: Academic Press.

Krishnapuram, R. and J.M. Keller. (1993) "A Possibilistic Approach to Clustering." IEEE Transactions on Fuzzy Systems, Vol.1, pp. 98-110.

Viattchenin, D.A. (2004) "A New Heuristic Algorithm of Fuzzy Clustering." Control and Cybernetics, Vol.33, pp. 323-340.

Viattchenin, D.A. (2006) "On the Inspection of Classification Results in the Fuzzy Clustering

Method Based on the Allotment Concept." In: Proceedings of 4th International Conference on Neural Networks and Artificial Intelligence. pp. 210-216.

Viattchenin, D.A. (2007) "A Direct Algorithm of Possibilistic Clustering with Partial Supervision." Journal of Automation, Mobile Robotics and Intelligent Systems, Vol.1, No.3, pp. 29-38.

Viattchenin, D.A., A. Damaratski and D. Novikau. (2009) "Relational Clustering of Heterogeneous Fuzzy Data." In: Developments in Fuzzy Clustering. Minsk: VEVER Publishing House, pp. 76-91.