Reliability of Clustering Methods-Based Uncertainty

Mika Sato-Ilic

Institute of Systems and Information Engineering, University of Tsukuba, Japan

Abstract

We define the reliability of the results of fuzzy clustering. In this case, we consider both the degree of belongingness of an object to a cluster and the classification situation. When considering both, we use t-norm defined in a statistical metric space. We also show that by using the asymmetric aggregation operator proposed by the author, it is possible to consider the difference in weights between the degree of belongingness and the classification situation.

1. Introduction

In today's world, where large amounts of complex data are being collected in a wide variety of fields, clustering methods are becoming increasingly important as they summarize data and extract its latent characteristics by grouping the data based on their similarities.

Among clustering methods, there is a methodology that takes into account the uncertainty of an object's cluster membership and attempts to extract the structure of real-world complex data with a smaller number of clusters.

In the past, classification methods based on statistics and probability theory were used as the basis, but recently clustering methods that more flexibly expand the solution space of the clustering partition matrix have been proposed. Fuzzy clustering is one of these types of clustering methods, and is a method for extracting fuzzy clusters defined based on fuzzy subsets, which are the basis of fuzzy logic. Fuzzy theory, along with neural networks and evolutionary computation, is a field of study based on soft computing and constitutes computational intelligence (CI) [1], [2], a branch of artificial intelligence.

The advantage of fuzzy clustering is that it allows for uncertainty in the membership of objects to clusters, making it possible to obtain classification results with excellent robustness and tractability for large-scale, complex data. However, on the other hand, as the classification results become more complex, they are difficult to interpret, and the reliability of the results is therefore an issue. In particular, with regard to the reliability of fuzzy clustering results, due to the definition of fuzzy subsets, the classification results cannot be measured with normal probability measures, and original validity functions are currently being developed.

Therefore, we define the reliability of fuzzy classification results by introducing the validity measure of fuzzy clustering into the fuzzy clustering results. An aggregation operator defined in a statistical measure space [3] is used to aggregate the clustering results and the validity function values. We also show that by applying the asymmetric aggregation operator developed by the author [4], it is possible to introduce weights that take into account the difference between the validity measure values and the clustering results. Furthermore, from the mathematical definition of these aggregation operators, we show that the proposed reliability measures make it possible to

remove noise in the data. Furthermore, the proposed method can be used for classification with learning in machine learning. This is because it utilizes the robustness of fuzzy clustering for large amounts of complex data, which solves the problem of the reliability of the training data. In other words, noise can be removed by converting objects from the training data to the solution space of fuzzy clustering. We demonstrate the effectiveness of our method through several numerical examples.

2. Reliability of Fuzzy Clustering

The results of fuzzy clustering are represented by the partition matrix $U = (u_{ik})$. Here, $u_{ik}, i = 1, \dots, n, k = 1, \dots, K$ is the degree of belongingness of an object i to a cluster k, and is assumed to satisfy the conditions $u_{ik} \in [0,1], \sum_{k=1}^{K} u_{ik} = 1$. Here, n is the number of objects, and K is the number of clusters, which is given in advance. Then, the reliability of an object i in cluster k is defined as in equations (1) and (2).

$$f_{ik} = \rho(u_{ik}, v_i). \qquad (1)$$
$$\tilde{f}_{ik} = g(u_{ik}, v_i). \qquad (2)$$

Here, v_i represents the classification status of an object i with respect to K clusters, and is a measure of validity for the fuzzy partition U shown in equation (3).

$$v_i = \sum_{k=1}^{K} u_{ik}^2, \quad v_i = 1 + \sum_{k=1}^{K} u_{ik} \log_K u_{ik}. \quad (3)$$

Furthermore, ρ in equation (1) is an aggregation operator which satisfies the following conditions:

$\rho: [0,\ 1] \times [0,\ 1] \to [0,\ 1] \qquad \forall a, b, c, d \in [0,\ 1]$
$0 \le \rho(a,b) \le 1,\ \ \rho(a,0) = \rho(0,a) = 0, \rho(a,1) = \rho(1,a) = a$ (Boundary conditions)
$a \le c,\ \ b \le d \Longrightarrow \rho(a,b) \le \rho(c,d)$ (Monotonicity)
$\rho(a,b) = \rho(b,a)$ (Symmetry)

T-norm function in the statistical metric space is a typical example that satisfies the above conditions of the aggregation operator ρ, and is generated by the generating function in equation (4).

$$\rho(x,y) = f^{[-1]}(f(x) + f(y))), \qquad (4)$$

where f is a continuous monotonically decreasing function as follows:

$$f: [0,1] \to [0,\infty],\ \ f(1) = 0, \qquad f^{[-1]}(z) = \begin{cases} f^{-1}(z), & z \in [0, f(0)) \\ 0, & z \in [f(0), \infty] \end{cases}.$$

Furthermore, g in equation (2) is a function that replaces the symmetric condition with an asymmetric condition among the conditions satisfied by the aggregation operator ρ, and its generating function is shown in equation (5).

$$g(x,y) = f^{[-1]}(f(x) + \varphi(x)f(y))), \qquad (5)$$

where φ is a continuous monotonically decreasing function, and satisfies the following conditions:

$$\varphi: [0,1] \to [0,\infty], \ \ \varphi(1) = 1.$$

That is, the asymmetric aggregation operator g satisfies the following conditions:

$g: [0,\ 1] \times [0,\ 1] \to [0,\ 1] \qquad \forall a, b, c, d \in [0,\ 1]$
$0 \le g(a,b) \le 1, \ \ g(a,0) = 0, \ \ g(a,1) = a$ (Boundary conditions)
$a \le c, \ \ b \le d \Longrightarrow g(a,b) \le g(c,d)$ (Monotonicity)
$g(a,b) \ne g(b,a), \ \ a \ne b$ (Asymmetry)

Therefore, in equation (2), $\tilde{f}_{ik} = g(u_{ik}, v_i) \ne g(v_i, u_{ik}), u_{ik} \ne v_i$.

3. Numerical Examples

Table 1 shows examples of t-norm for Equation (4). Figure 1 shows the difference in reliability defined by Equation (1) due to different t-norms. The horizontal axis represents fuzzy cluster 1, and the vertical axis represents fuzzy cluster 2. The solid lines show the membership degrees obtained as a result of fuzzy clustering, and the numbers 1 to 11 represent object numbers. The dashed lines show the reliability values for different t-norms shown in Table 1. From this figure, we can see that excluding the minimum and bounded product, if we obtain an uncertainty clustering result, then the degree of reliability also decreases. Also, greater uncertainty results make larger the difference in the degree of reliability depending on the differences of the t-norm. These findings demonstrate the validity of the proposed reliability measure for fuzzy clustering results. In addition, the minimum and bounded product are not adaptable for the degree of reliability, because the minimum is not Archimedean and the bounded product tends to have the same values.

In Figure 2, (a) shows the reliability using the asymmetric aggregation operator $g(x,y) = xy^{(2-x)^2}$ generated using the generating function for algebraic product f and $\varphi(x) = (2-x)^2$ in equation (2). (b) shows the reliability using the asymmetric aggregation operator $g(x,y) = x^3y/(1-y+x^2y)$ generated using the generating function for Hamacher product f and $\varphi(x) = 1/x^3$, excluding the cases where the membership is (1,0) or (0,1). These figures show that the asymmetry of the asymmetric aggregation operator is equivalent to changing the weight related to the validity of the reliability, and that it can be used depending on whether the fuzzy clustering results or the classification situation are considered more important.

Table 2 shows the frequency of device selection by students for each type of assignment in English classes. In this table, the assignments are divided into "L: Listening; R: Reading; S: Speaking; W: Writing," and the number after the letter indicates the number of words. For example, "L1" indicates a single-word listening assignment. Figure 3 shows the result of a principal component analysis of the data in Table 2. Figure 3 reveals that students select device types depending on whether the number of words is large or small, and whether the assignment is "Listening/Speaking" or "Reading/Writing." Based on the results in Figure 3, fuzzy clustering was performed with four

clusters. Figure 4 shows the membership and reliability results for Cluster 1 and Cluster 2. Figure 5 shows the results of other cluster combinations. From these figures, we can see the same four clusters obtained. Cluster 1 is “Listening and Speaking large words”, Cluster 2 is “Writing and Reading large words”, Cluster 3 is “Listening and Speaking small words”, and Cluster 4 is “Writing and Reading small words”. Also, if the objects are an uncertain classification structure, then the degree of reliability of these objects is going to be significantly smaller when compared with the objects that are clearly classified into a cluster. These findings suggest that reliability may be able to remove noise from the initially obtained membership results.

Table 1　Examples of t-norm

t-norm	$t(x,y)$	$f(x)$
Minimum	$\min\{x,y\}$	(*)
Hamacher Prod.	$\dfrac{xy}{x+y-xy}$	$\dfrac{1-x}{x}$
Algebraic Prod.	xy	$-\log x$
Einstein Prod.	$\dfrac{xy}{1+(1-x)(1-y)}$	$\log\dfrac{2-x}{x}$
Bounded Prod.	$\max\{0,x+y-1\}$	$1-x$

(*)min does not have a generating function (not Archimedian)

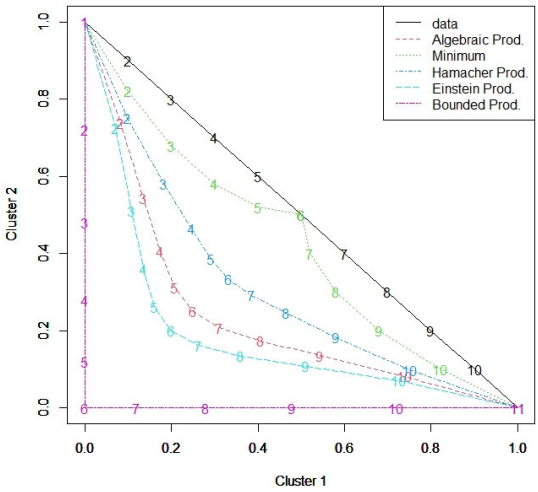
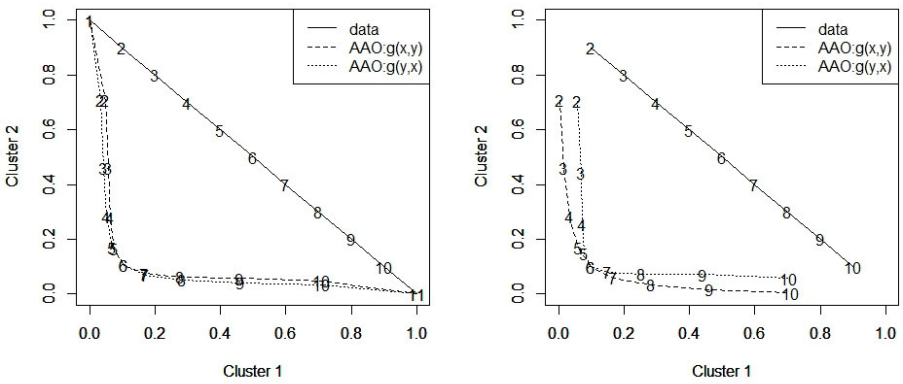


Figure 1 Reliability of fuzzy clustering depending on t-norm



(a) Generating function is algebraic product　(b) Generating function is Hamacher product

Figure 2 Asymmetric reliability

Table 2　Frequency of device selection by task type

Situations	Desktop PC	Laptop PC	Tablet	E-reader	Smartphone	Paper
L1	29	60	31	12	76	15
L20	25	61	31	12	78	14
L150	35	79	29	11	56	9
L500	40	84	27	9	41	8
L2000	42	85	21	9	28	10
L50000	41	80	19	6	27	13
R1	34	59	34	17	69	40
R20	32	62	34	18	67	37
R150	36	74	35	18	50	36
R500	41	87	34	16	24	40
R2000	38	85	25	12	18	47
R50000	35	75	22	8	13	47
S1	28	51	22	7	67	12
S20	29	53	20	6	65	10
S150	31	69	23	8	48	13
S500	31	69	23	6	37	12
S2000	32	69	17	3	30	12
S50000	32	66	15	2	28	12
W1	33	65	27	8	48	44
W20	37	71	27	7	44	42
W150	44	85	19	6	23	34
W500	45	86	14	4	12	27
W2000	45	87	8	3	6	27
W50000	45	83	6	1	5	26

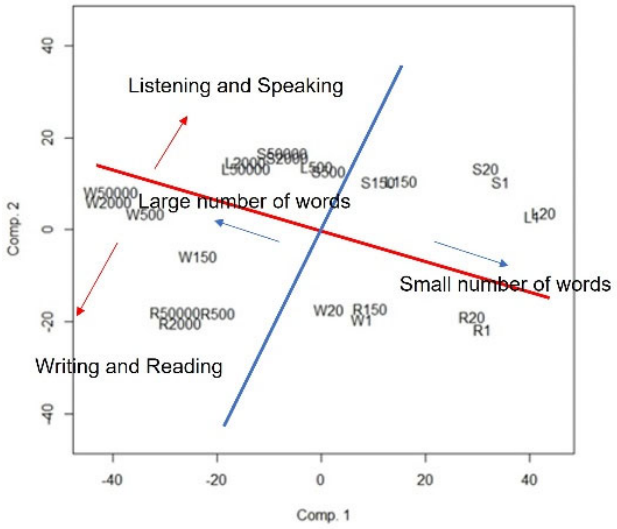


Figure 3 Result of principal component analysis

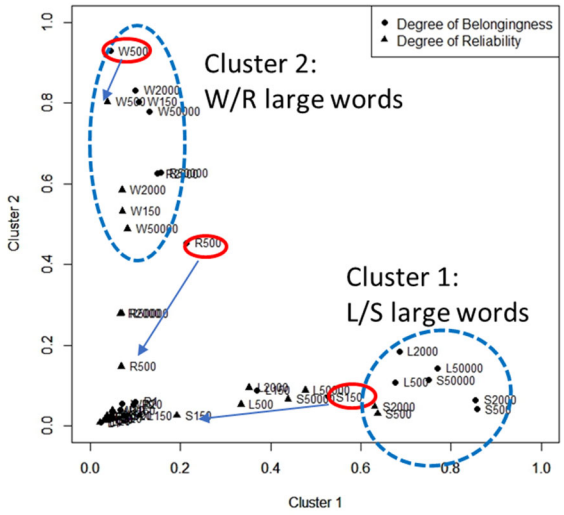


Figure 4 Degree of belongingness of device selection data and results of degree of reliability with respect to clusters 1 and 2

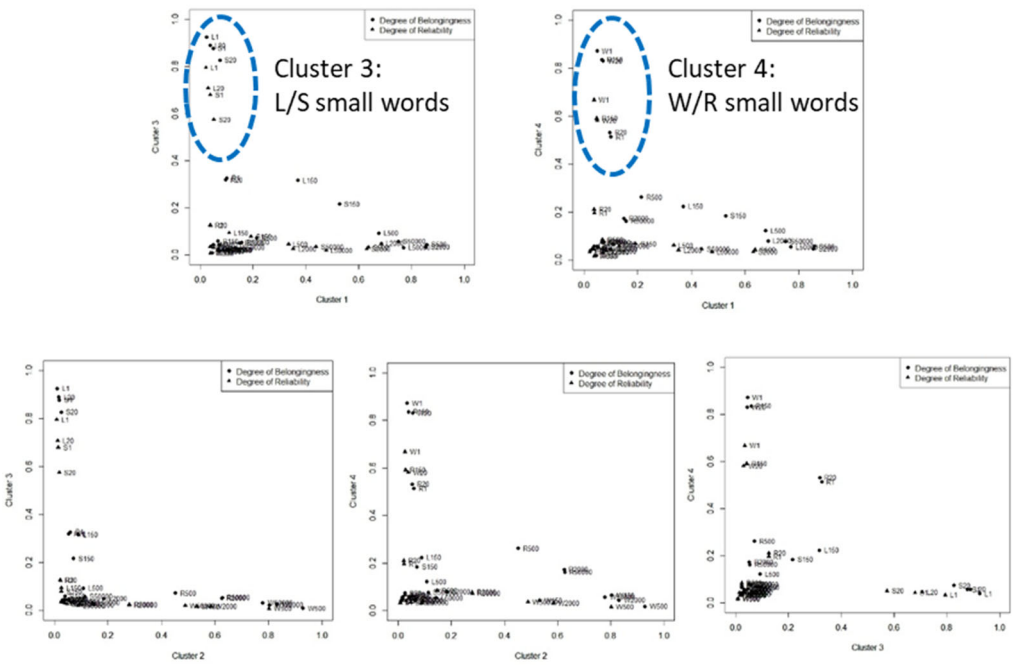


Figure 5 Degree of belongingness of device selection data and results of degree of reliability with respect to other cluster combinations

4. Conclusion

This paper defines the reliability of fuzzy clustering results considering both the degree of belongingness and classification status of objects to clusters. To consider both, we use aggregation operators. Especially, by using an asymmetric aggregation operator, we can show the differences in weights between the degree of belongingness and classification status. For the future study, other validity measures of fuzzy clustering, such as min-max operation, weighted intraclass and interclass sum of squared deviations, hypervolume, or partition density measures, will be used for obtaining classification status of objects to clusters in the proposed reliability of fuzzy clustering results. Also, we believe this degree of reliability can be used for the appropriate selection of t-norms.

Reference

[1] J.C. Bezdek (1994). What is a Computational Intelligence?. In J.M. Zurada, R.J. Marks II, C.J. Robinson (Eds.), Computational Intelligence: Imitating Life, 1–12, IEEE Press

[2] IEEE Computational Intelligence Society: https://cis.ieee.org/about/what-is-ci

[3] K. Menger (1942). Statistical Metrics, Proc. Nat. Acad. Sci. USA, 28, 535-537

[4] M. Sato-Ilic, Y. Sato (2000). Asymmetric Aggregation Operator and its Application to Fuzzy Clustering Model, Computational Statistics & Data Analysis, 32, 379-394

[5] M. Sato-Ilic (2024). On Reliability of Fuzzy Clustering, Proceedings of the 43rd Annual Meeting of Japanese Classification Society, 45-48 (in Japanese)