Mathematics and Statistics 8(5): 559-565, 2020
DOI: 10.13189/ms.2020.080509 http://www.hrpub.org

Integration of Cluster Centers and Gaussian Distributions in Fuzzy C-Means for the Construction of Trapezoidal Membership Function

Siti Hajar Khairuddin*, Mohd Hilmi Hasan, Manzoor Ahmed Hashmani

Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Perak, Malaysia

Received April 13, 2020; Revised July 14, 2020; Accepted August 5, 2020

Cite This Paper in the following Citation Styles

(a): *[1] Siti Hajar Khairuddin, Mohd Hilmi Hasan, Manzoor Ahmed Hashmani , "Integration of Cluster Centers and Gaussian Distributions in Fuzzy C-Means for the Construction of Trapezoidal Membership Function," Mathematics and Statistics, Vol. 8, No. 5, pp. 559 - 565, 2020. DOI: 10.13189/ms.2020.080509.*

(b): *Siti Hajar Khairuddin, Mohd Hilmi Hasan, Manzoor Ahmed Hashmani (2020). Integration of Cluster Centers and Gaussian Distributions in Fuzzy C-Means for the Construction of Trapezoidal Membership Function. Mathematics and Statistics, 8(5), 559 - 565. DOI: 10.13189/ms.2020.080509.*

Abstract Fuzzy C-Means (FCM) is one of the mostly used techniques for fuzzy clustering and proven to be robust and more efficient based on various applications. Image segmentation, stock market and web analytics are examples of popular applications which use FCM. One limitation of FCM is that it only produces Gaussian membership function (MF). The literature shows that different types of membership functions may perform better than other types based on the data used. This means that, by only having Gaussian membership function as an option, it limits the capability of fuzzy systems to produce accurate outcomes. Hence, this paper presents a method to generate another popular shape of MF, the trapezoidal shape (trapMF) from FCM to allow more flexibility to FCM in producing outputs. The construction of trapMF is using mathematical theory of Gaussian distributions, confidence interval and inflection points. The cluster centers or mean (μ) and standard deviation (σ) from the Gaussian output are fully used to determine four trapezoidal parameters; lower limit *a*, upper limit *d*, lower support limit *b*, and upper support limit *c* with the assistance of function trapmf() in Matlab fuzzy toolbox. The result shows that the mathematical theory of Gaussian distributions can be applied to generate trapMF from FCM.

Keywords Fuzzy C Means, Gaussian Distribution, Normal Distribution Membership Function, Trapezoidal MF

1. Introduction

Fuzzy logic is an idea whereby we apply the uncertainties in the real world to be applied in the computing world which represents the degree of truth. It contradicts with the crisp value or Boolean value (0 or 1) to produce a certain result, which is not realistic [1][2]. The idea of fuzzy logic was introduced by Dr Lotfi Zadeh when he was working in natural language which cannot be easily translated into absolute terms of true or false [3]. An application of fuzzy logic can be found in fuzzy inference system (FIS). Fuzzification is a component in an FIS where an input variable is compared to a membership function (MF) to obtain the membership degree [4]. The membership degree will go through the fuzzy rule engine for processing, such as decision making. Membership degree of a fuzzifier can be constructed by two methods: expert opinion and generated via data [5]. One of the common methods for MF construction from data is through clustering.

Fuzzy clustering is a technique to handle unlabelled data, which may contain outliers and unusual patterns. Thus, membership functions can provide the possibility of one data point to belong to other groups or clusters [6]. The clusters of data are generated by a possibility distribution

or collected from various resources. The measurement used in most clustering algorithms to determine the cluster centres is Euclidian distance [7]. Fuzzy C-Means (FCM) is one of the mostly used techniques for fuzzy clustering [8]. Based on various applications such as web usage mining, stock market, web analytics and image segmentation, FCM is proven to be more efficient, robust and reliable by its performance [9]. However, the resultant MF of FCM is of only Gaussian shape due to a straightforward nature of clusters to be distributed [10]. There are other regularly used parameterized MFs such as triangular and trapezoidal MFs. These MFs are used in specific cases such as antenna positioning fuzzy controller [11] and crime prevention analysis [12]. Thus, FCM should also have a capability to produce linear MFs. The construction of Gaussian MFs is straightforward and discussed in [13]. The construction of trapezoidal MFs is not quite straightforward and a method of convex hull is proposed in [12][13]. However, the implementation is unclear and it depends on specific cases [16].

In this paper, we present a method to produce trapezoidal MFs based on the integration of cluster centres produced by FCM and mathematical theory of Gaussian distributions. This paper is organized as follows: In section 2, literature review on trapezoidal MF and Gaussian distribution is introduced and an approximation of trapezoidal MFs is explained. The result and testing are in Section 3, and Section 4 presents the conclusion.

2. Literature Review

Trapezoidal MF

According to [17], with experience, one can decide which shape of MF is good for certain application and cases under consideration. This is where the degrees of freedom is offered in the fuzzy system environment since the MFs can be of any shape and form as long as it could map the given datasets with the desirable membership degrees. It also depends entirely on the size and type of the problem. The MF shapes are not the only concern as setting up the interval and the numbers of MFs are considered important too [17]. In addition, trial and error method is often used to determine the shape of MF. However, trapezoidal MF (trapMF), which represents fuzzy intervals, is proven to be easy to implement with fast computation [17]. In most practical applications, trapMF functions work well since they use linear interpolation to get both endpoints of the interval. The theoretical explanation which proves the practicality of trapMFs is discussed in [18]. An intuitive explanation on how trapMF is functioning well is explained in simple terms in [19]. In short, a trapMF (Figure 1) is defined by a lower limit a, an upper limit d, a lower support limit b, and an upper support limit c, where $a < b < c < d$ [18]. Compactly, the mathematical representation of trapMF is as follows:

$$f(x;a,b,c,d) = \max\left(\min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), o\right) \quad (1)$$

In a simple explanation, as an example, for a property such as “small”, different values namely x are assigned, along with their degrees, $\mu(x)$. The main motivation in fuzzy implementation is to ensure the value x and x' are close, along with their corresponding MFs $\mu(x)$ and $\mu(x')$ which should be close too [19].

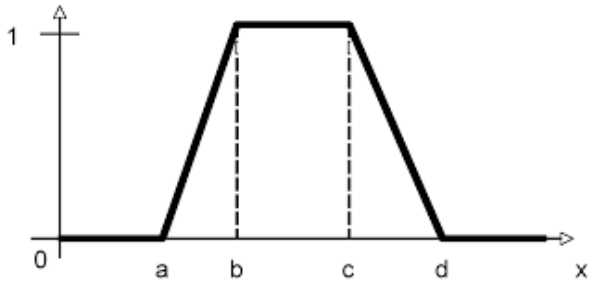

Figure 1. Trapezoidal MF

Gaussian Distribution

Since trapMF in this paper is constructed via FCM, the output of the data clustering is in the form of Gaussian MF. Hence, it is natural to use Gaussian distribution to convert the Gaussian MF into trapMF. Gaussian distribution is also called normal distribution which states that a random variable is normally distributed. A normal distribution is informally called the bell curve. The function to calculate the probability of a random variable to be within a particular range of values, instead of taking on any one value is called the probability density function, shown in eq. (2).

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

where μ is the mean, σ is the standard deviation and σ^2 is the variances. Since the mean and standard deviation are provided from the output of Gaussian MF, it is possible to mathematically construct the trapMF by using the Gaussian distribution.

In a standard normal distribution, μ = 0 and σ = 1, and it is described by the probability density function (3),

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad (3)$$

whereby the factor $1 / \sqrt{2\pi}$ in this expression makes sure that the area under curve is equal to 1. Since the curve is symmetric, the inflection points are $x = +1$ and $x = -1$. The standard normal probability density function plot is as Figure 2.

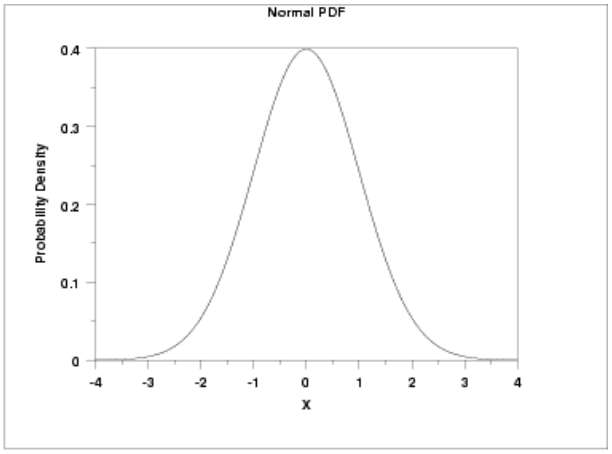

Figure 2. Probability Density Function

Based on the standard normal distribution in eq. (3), the value of the mean (cluster centers) and standard deviation obtained from the Gaussian MF will give information to the construction of trapMF based on the area under curve. Through the theory of inflection point, we can use the interval estimate such as the confidence interval and inflection points to approximate the range of lower and upper limits *a*, *b*, *c* and *d* of the trapezoidal shape.

Confidence Interval

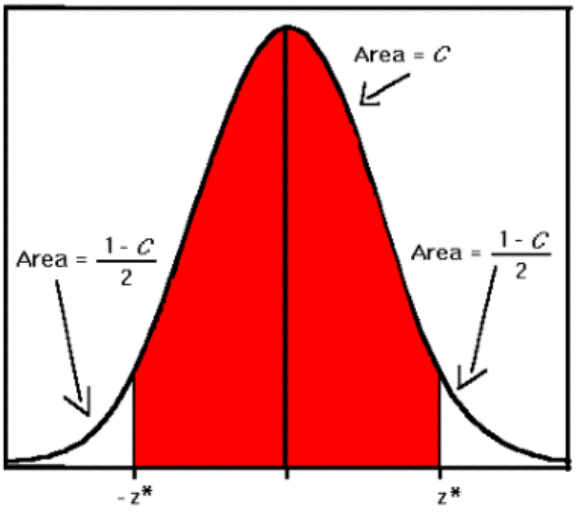

Figure 3. Confidence Interval of 95%

A confidence interval is a range of values where it is commonly known that the true value lies in [21]. This can also be obtained by a known mean and standard deviation (sigma) which makes the range to be helpful to approximate the trapMF data distributions. It is suitable to be used to estimate the range calculated from a given dataset [22] The common choices for the confidence level, *C* are 0.90, 0.95 and 0.99 which correspond to a normal Gaussian curve area percentage. Calculations for both left half and the right half of the curve will be the same since the Gaussian curve is symmetrical, which naturally will produce a symmetrical trapezoid shape. As shown in Figure 3, each tail of the curve has the area which is equal to (1-*C*) / 2. The area in each tail is equal to 0.05/2 = 0.025 for a 95% confidence interval [22].

The value z^* in Figure 3 is representing the inflection points on the standard normal density Gaussian curve which shows that the probability of observing a value greater than z^* which equals to *p* is known as the upper *p* critical value of the standard normal distribution. For a confidence interval with level *C*, the value *p* is equal to (1-*C*)/2. The interval is (-1.96, 1.96), since 95% of the area under the curve falls within this interval [22].

Inflection Points Approximations for TrapMF

Inflection points are where the curve turns inwards or concave where for a Gaussian shape, it will concave near the peak in a downward manner [23]. The inflection points need to be obtained from the probability density function of the Gaussian in order to locate the x-axis points. According to [23], inflection points are normally at ± σ, ±α and $\alpha\sqrt{2}$. These points are applied for symmetrical and normal Gaussian shape.

The probability density function (PDA) for a normally distributed random variable with a known mean μ and standard deviation is in eq. (4).

$$f(x) = 1/(\sigma\sqrt{(2\pi)})\exp[-(x-\mu)^2/(2\sigma^2)] \quad (4)$$

The notation $\exp[y] = e^y$ is used where e is a constant of approximately by 2.71828 [24]. The first derivative of the PDA is found by getting the derivative for e^x and the chain rule is applied in eq. (5).

$$f'(x) = -(x-\mu)/(\sigma^3\sqrt{(2\pi)})\exp[-(x-\mu)^2/(2\sigma^2)]$$
$$= -(x-\mu) f(x)/\sigma^2 \quad (5)$$

The second derivative of the PDA is calculated by using the product rule in eq. (6):

$$f''(x) = -f(x)/\sigma^2 - (x-\mu) f'(x)/\sigma^2 \quad (6)$$

The simplified expression is in eq. (7).

$$f''(x) = -f(x)/\sigma^2 + (x-\mu)^2 f(x)/(\sigma^4) \quad (7)$$

The expression in (7) is set to zero to solve for x. Since f(x) is a non zero function [24], both sides of the equation can be divided by the function in eq. (8):

$$0 = -1/\sigma^2 + (x-\mu)^2/\sigma^4 \quad (8)$$

Both sides can be multiplied by σ^4 to eliminate the fractions [24] as shown in eq. (9).

$$0 = -\sigma^2 + (x-\mu)^2 \quad (9)$$

In order to solve x (inflection point for each side), by using $\sigma^2 = (x-\mu)^2$, calculate the square root for both sides, to produce eq. (10).

$$\pm\sigma = x - \mu \quad (10)$$

Based on the equation, the inflection points occurred

when $x = \mu \pm \sigma$ [24]. It is located one standard deviation above the mean and one standard deviation below the mean. Hence, the cluster center obtained from FCM clustering is used in eq. (10) and it is subtracted from the standard deviation. It is applied for both left and right side of the curve. To get the inflection points of lower limit a, and upper limit d, the theory of confidence interval (-1.96, 1.96) is used whereby an approximation of $x = \mu \pm 2\sigma$ suits the result for the datasets in testing process.

3. Result and Testing

For the purpose of testing the theory, a dataset which contains the response time of a web service is used in order to generate the GaussianMF from FCM. The dataset consists of 6145 points and is milliseconds (ml) in unit. The dataset is one of the attributes to access the quality of web service and it is obtained from online resources. FCM will generate two outputs, mean and standard deviation, which will be used for trapMF approximation. Figure 4 shows the result from FCM. The number of clusters (three in this particular sample) is determined by using Clustering Validity Index (CVI) [5].

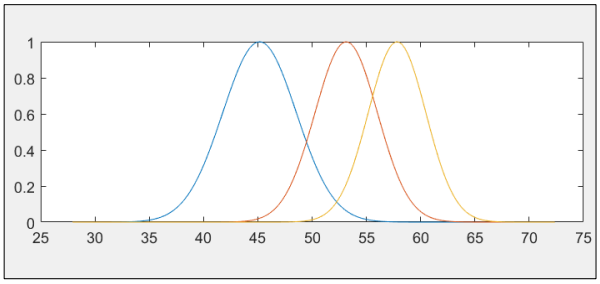

Figure 4. Gaussian MF

By using the cluster centers (μ) and sigma (σ) of the Gaussian output, trapMF approximation is performed. Both values are maintained while the inflection points are tested in trial and error manner based on eq. (10). Figure 5 shows the trapMF generated from the approximations for the respective GaussianMF.

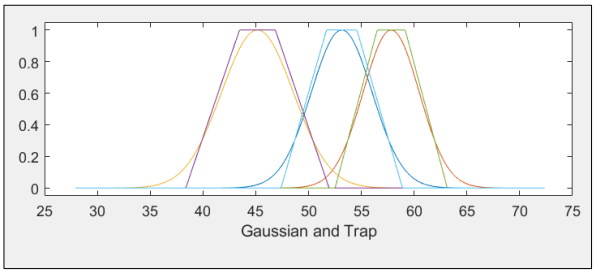

Figure 5. TrapMF and GaussianMF

Figure 6 shows the trapMF after being separated by its respective GaussianMF. The trapMF will be compared with trapMF generated by toolbox in Matlab to validate its parameters.

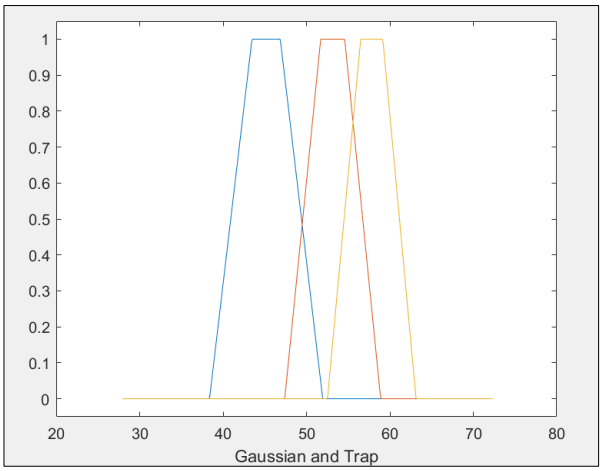

Figure 6. TrapMF

In Figure 7, fuzzy toolbox in Matlab is used to validate the trapMF generated from the mathematical calculation with the trapMF generated by toolbox. A function called evalmf() is used to evaluate the result.

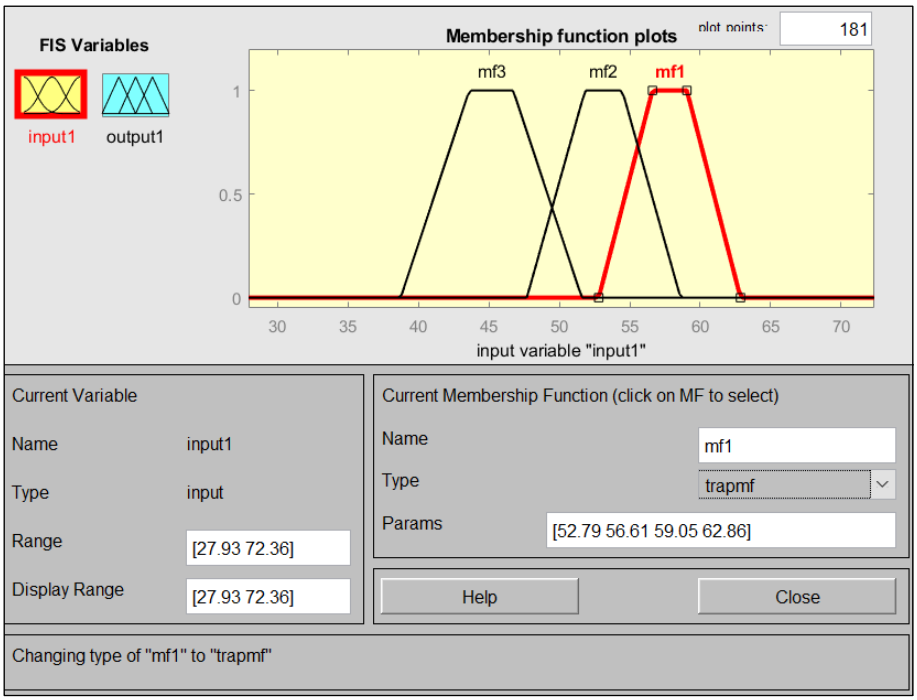

Figure 7. Matlab implementation of trapMF

Table 1 shows the cluster centers and sigma for the chosen dataset. From the obtained results, approximation from GaussianMF and TrapMF is conducted and it produces the parameters *a*, *b*, *c* and *d* as in Table 2. To get the parameters, a function to produce trapezoidal membership functions called trapmf() in Matlab Fuzzy toolbox is used. To use trapmf(), the membership function range is obtained from FCM output and applied in the toolbox. Next, set the type to trapmf(). The availability of the toolbox to generate and edit the membership functions as well as to design fuzzy inference system may help users to apply fuzzy logic. However, mathematical solution to generate membership function is still important since dealing with a toolbox usually involves software compatibility issue and is not publicly available.

Table 1. Mean and Sigma for GaussianMF

Cluster	Mean (μ)	Sigma (σ)
Cluster 1	2.6558	57.8258
Cluster 2	2.8754	53.1434
Cluster 3	3.3927	45.1600

Table 2. TrapMF Parameters

Cluster	*a*	*b*	*c*	*d*
Cluster 1	38.3746	43.4637	46.8564	51.9455
Cluster 2	47.3926	51.7057	54.5812	58.8943
Cluster 3	52.5141	56.4979	59.1538	63.1377

Next, to test the significance between the manually generated and Matlab toolbox-generated trapMF, t-test is performed. The p-value, which is the probability that the results from the dataset are occurred by chance will be the ratio that will determine the validity of the result. P <0.05 means the data have statistically significant difference [25]. Our target is to validate whether the output produced by manually generated trapMF is not significantly different from the toolbox-generated trapMF. Table 3, 4 and 5 show the t-test result from the web response time dataset for each cluster from the FCM outputs. Based on the p values, it proves that the results are valid and statistically significant.

Table 3. T-test for cluster 1

t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	*Variable 2*
Mean	0.17954918	0.190601
Variance	0.11108443	0.116411
Observations	445	445
Hypothesized Mean Difference	0	
df	888	
t Stat	-0.48880072	
P(T<=t) one-tail	0.31255174	
t Critical one-tail	1.64657139	
P(T<=t) two-tail	0.62510348	
t Critical two-tail	1.96263904	

Table 4. T-test for cluster 2

t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	*Variable 2*
Mean	0.1405656	0.149223
Variance	0.0924348	0.09733
Observations	445	445
Hypothesized Mean Difference	0	
df	887	
t Stat	-0.41925	
P(T<=t) one-tail	0.3375675	
t Critical one-tail	1.6465733	
P(T<=t) two-tail	0.6751351	
t Critical two-tail	1.9626421	

Table 5. T-test for cluster 3

t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	*Variable 2*
Mean	0.152251	0.161556
Variance	0.098316	0.103371
Observations	445	445
Hypothesized Mean Difference	0	
df	887	
t Stat	-0.43705	
P(T<=t) one-tail	0.331092	
t Critical one-tail	1.646573	
P(T<=t) two-tail	0.662184	
t Critical two-tail	1.962642	

4. Conclusions

In this paper, we present a method to generate trapezoidal MFs based on the integration of cluster centers produced by FCM and mathematical theory of Gaussian distributions. MF is developed by using FCM clustering method in Matlab environment. The mean and sigma of the Gaussian output is then used to mathematically construct the trapMF. In overall, the proposed method can provide more flexibility to FCM when it allows the generation of other membership function types. For future works, the proposed method may be further explored to generate trapezoidal fuzzy type-2 membership functions.

Acknowledgements

This research is an ongoing research supported by Fundamental Research Grant Scheme (FRGS/1/2018/ICT02/UTP/02/1); a grant funded by the Ministry of Education, Malaysia.

REFERENCES

[1] Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

[2] Mendel, J. M., John, R. I., & Liu, F. (2006). Interval Type-2 Fuzzy Logic Systems Made Simple. IEEE Transactions on Fuzzy Systems, 14(6), 808–821. https://doi.org/10.1109/TFUZZ.2006.879986

[3] Li, Jiamin & W. Lewis, Harold. (2016). Fuzzy Clustering Algorithms — Review of the Applications. 282-288. 10.1109/SmartCloud.2016.14.

[4] Jang, J.-R. (1993). ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics, 23(3), 665–685. https://doi.org/10.1109/21.256541Kaymak, U. and Setnes, M. (2000). Extended Fuzzy Clustering Algorithm. ERIM Report Series Research in Management. 1-23.

[5] M.H. Hasan, J. Jaafar, M.F. Hassan, (2016). Fuzzy C-Means and Two Clusters' Centers Method for Generating Interval Type-2 Membership Function, International Conference on Computer and Information Sciences (ICCOINS) 2016.

[6] Li, Jiamin & W. Lewis, Harold. (2016). Fuzzy Clustering Algorithms — Review of the Applications. 282-288. 10.1109/SmartCloud.2016.14.

[7] Dunn, J. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact, Well-Separated Cluster. Journal of Cybernetics.3(3). 32-57.

[8] Kaymak, U. and Setnes, M. (2000). Extended Fuzzy Clustering Algorithm. ERIM Report Series Research in Management. 1-23.

[9] Singh, T., & Mahajan, M. (2014). Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm. International Journal of Advanced Research in Computer Science and Software Engineering, 4(5). Retrieved from https://pdfs.semanticscholar.org/3720/5c8fe390d36bde67a2e0f614d5ce8bba829b.pdf

[10] Castillo, O. and Melin, P., (2008). Design of Intelligent Systems with Interval Type-2 Fuzzy Logic. Type-2 Fuzzy Logic: Theory and Applications - Studies in Fuzziness and Soft Computing, 223, pp. 53-76.

[11] Kalist, V., Ganesan, P., Sathish, B.S., Jenitha, J.M.M., (2015). Possiblistic-Fuzzy C-Means Clustering Approach for the Segmentation of Satellite Images in HSL Color Space. Procedia Computer Science, 57, pp.49-56.

[12] Wang, L. and Wang, J. (2012). Feature Weighting fuzzy clustering integrating rough sets and shadowed sets. International Journal of Pattern Recognition and Artificial Intelligence, 26(4).

[13] Rajen Bhatt and M.Gopal (2006). "Neuro-fuzzy decision trees". International Journal of Neural Systems.vol. 16, no.1,

pp. 63-78.

[14] M. Sugeno and T. Yasukawa (1993). “A fuzzy-logic based approach to qualitative modeling”, IEEE Transactions on Fuzzy Systems, vol. 1, pp.6-31.

[15] M.R. Emami, I.B. Turksen, and A.A. Goldenberg (1998). Development of a systematic methodology of fuzzy logic modeling. IEEE Transactions on Fuzzy Systems, vol. 6, no. 3, pp. 346-361.

[16] Bhatt, R.B., Narayanan, S.J., Paramasivam, I. and Khalid, M., (2012). Approximating Fuzzy Membership Functions from Clustered Raw Data. 2012 Annual IEEE India Conference (INDICON).

[17] Sadollah, A. (2018, October 31). Introductory Chapter: Which Membership Function is Appropriate in Fuzzy System? Retrieved from https://www.intechopen.com/books/fuzzy-logic-based-in-optimization-methods-and-control-systems-and-its-applications/introductory-chapter-which-membership-function-is-appropriate-in-fuzzy-system-

[18] Barua, A & Mudunuri, L.S. & Kosheleva, Olga. (2014). Why Trapezoidal and Triangular Membership Functions Work So Well: Towards a Theoretical Explanation. Journal of Uncertain Systems. 8. 164-168.

[19] Kreinovich,, V., Kosheleva, O., & Shabazova, S. (2018). Why Triangular and Trapezoid Membership Functions: A Simple Explanation. Center of Excellence). Retrieved March 8, 2019, from http://www.cs.utep.edu/vladik/2018/tr18-59.pdf

[20] Reyna Vargas, M. E. (2018). Fuzzy Analytical Hierarchy Process Approach for Multicriteria Decision-Making with an Application to developing an ‘Urban Greenness Index’. (Masters Thesis). University of Toronto. Retrieved from http://hdl.handle.net/1807/91594

[21] P.A., Wasserman, S.S., and Levine, M.M. (1992), "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich," Journal of the American Medical Association, 268, 1578-1580.

[22] Cox D.R., Hinkley D.V. (1974) Theoretical Statistics, Chapman & Hall, p49, p209

[23] DeBruyne, D., & Sorensen, L. (2018). Quantum Mechanics I. Walter de Gruyter GmbH.

[24] Taylor, Courtney. (2019, April 28). How to Find the Inflection Points of a Normal Distribution. Retrieved from https://www.thoughtco.com/inflection-points-of-a-normal-distribution-3126446

[25] T Test (Student's T-Test): Definition and Examples. (n.d.). Retrieved fromhttps://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/

[26] Nazirah Ramli, Siti Musleha Ab Mutalib, Daud Mohamad, Fuzzy Time Series Forecasting Model based on Centre of Gravity Similarity Measure, Journal of Computer Science & Computational Mathematics, Vol. 8, No. 4, pp. 121-124, 2018..