IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 21, 2020, accepted June 25, 2020, date of publication June 29, 2020, date of current version July 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3005666

Data Clustering Method Using Efficient Fuzzifier Values Derivation

JAEHYUK CHO AND WONHEE JOO

Department of Electronic Engineering, Soongsil University, Seoul 06978, South Korea

Corresponding author: Jaehyuk Cho (wogur900@gmail.com)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by the Korea Government, Ministry of Science and ICT (MSIT) (Implementation of Verification Platform for ICT Based Environmental Monitoring Sensor) under Grant 2019-0-00136.

ABSTRACT The Type-2 fuzzy set (T2 FS) is widely used for efficient control uncertainties, such as noise sensitivity in the fuzzy set. In addition, unsupervised machine learning requires a clustering parameter value in advance, and may affect clustering performance according to prior information such as the number and size of clusters. In this case, the fuzzifier value m to be applied is the most important factor in improving the accuracy of data. Therefore, in this paper, we intend to perform clustering to automatically acquire the determination of m_1 and m_2 values that depended on existing repeated experiments. To this end, in order to increase efficiency on deriving appropriate fuzzifier value, we used the Interval type-2 possibilistic fuzzy C-means (IT2PFCM), clustering method to classify a given pattern. In Efficient IT2PFCM method, used for clustering, we propose an algorithm that derives suitable fuzzifier values for each data. These values also extract information from each data point through the histogram approach and Gaussian Curve Fitting method. Using the extracted information, two adaptive fuzzifier value m_1 and m_2 are determined. Obtained values apply the new lowest and highest membership values. In addition, it is possible to form an appropriate fuzzy area on each cluster by only taking advantage of the characteristics of IT2PFCM, which reduces uncertainty. This doesn't only improve the accuracy of clustering of measured sensor data, but can also be used without additional procedures such as data labeling or the provision of prior information. It is also efficient at monitoring numerous sensors, managing and verifying sensor data collected in real time such as smart cities. Eventually, in this study, the proposed method is to improve IT2PFCM performance on accurate and quick clustering of large amount of complex data such as Internet of Things (IoT).

INDEX TERMS Fuzzifier value determining, sensor data clustering, fuzzy C-means, histogram approach, interval type-2 PFCM.

I. INTRODUCTION

Clustering is the process of grouping similar entities together, taking specific predefined features or attributes into consideration. In machine learning, one of clustering techniques using unsupervised learning, inferences are drawn from datasets consisting of input data without labelled responses. To establish the inaccurate and ambiguous of the fuzzy sets, concept of membership values is introduced. The membership values denote the degree with which an element x from the universe of discourse belongs to a particular set, where the membership value varies from 0 (not belonging to the set) to 1 (complete membership in the set). In other words, clustering can be thought of as two types, hard clustering and soft clustering. In hard clustering, the data points are divided into distinct sets, that is, a single data point belongs to only one cluster, whereas in soft clustering, data points have a fuzzy membership in a cluster, that is, a particular data point belongs to more than one cluster, containing different membership value. While traditional hard clustering works for physical systems, fuzzy clustering, a kind of soft cluster, is preferred for realistic human-centered systems.

Various algorithms have been previously introduced to solve unsupervised clustering problems of fuzzy sets. Many studies have been conducted on fuzzy clustering to classify patterns and fuzzy C-means (FCM) algorithm has been used most frequently [1]. FCM uses the concept of a fuzzifier m which is used to determine the membership value of a

The associate editor coordinating the review of this manuscript and approving it for publication was Jonghoon Kim.

pattern X_k belonging to a particular cluster with cluster prototype, here the cluster center, v_i where $k = 1, 2...\text{n}$ and $i = 1, 2...\text{c}$, where n is the number of patterns and c is the number of clusters. FCM requires the knowledge of the initial number of desired clusters and the membership value is decided by the relative distance between the pattern X_k and the cluster center V_i. However, one of the major drawbacks of using FCM is its noise sensitivity and constrained memberships. In order to solve problems of FCM method, PCM uses a parameter given by whose value is estimated from the dataset itself. PCM applies the possibilistic approach which simply means that the membership value of a point in a class represents the typicality of the point in the class, or the possibility of the pattern X_k belonging to the class with cluster prototype V_i where $k = 1, 2...\text{n}$ and $i = 1, 2...\text{c}$. Since, the noise points are comparatively less typical, while using typicality in PCM algorithm, the noise sensitivity is considerably reduced [2], [3]. However, the PCM algorithm also has a problem that the clustering result is sensitively reacted according to the initial parameter value [4]. To solve this problem, PFCM algorithm generated both the memberships and possibilities simultaneously and solved the problem of noise sensitivity as seen in FCM and the coincident clusters as experienced in PCM. FCM and PCM, where, the constraint on typicality values (or the constraint of row-sum $= 1$) is relaxed but the column constraints on membership values is retained. PFCM uses the fuzzifier that is denoted by m, which determines the membership values, and the bandwidth parameter that is used to evaluate the typicality values [5]. PFCM further uses constants a and b that define the relative importance of fuzzy membership and typicality values in the objective function. Since PFCM utilizes more number of parameters to decide on the optimal solution for clustering, it provides an increased degree of freedom and hence renders better results as compared to the research stated above. However, when we consider fuzzy sets and different parameters in a particular algorithm, we come across the possibility of the fuzziness of these parameters. In this paper, we account for the fuzziness in the possible value of the fuzzifier value m and the bandwidth parameter and generate a Footprint of uncertainty (FOU) for both by taking an interval of fuzziness for m, that is, considering the possibility of m lying in the interval m_1 and m_2, and an interval of fuzziness. The existing research has been conducted to measure the optimum range according to the upper and lower bounds of the fuzzifier value through several repeated experiments [6]. Although these studies are ongoing, the same fuzzy constant range cannot be applied to every data [7]. As the needs on developing new method to adaptively determining the fuzzifier value for different kinds of data are growing, this paper proposes a method using a histogram based on the Interval type-2 possibilistic Fuzzy C-means (IT2 PFCM) clustering method.

Section 2 introduces the concept of recent research trends, fuzzy values, and decisions, and section 3 describes the IT2 PFCM algorithm as a formula. Section 4 uses HISTOGRAM to determine the FUZZIFIER VALUE and presents the formula and FIGURE related to it. Section 5 actually performs test comparison to apply the sensor data to the proposed algorithm. Finally, section 6, "Conclusion," presents a contribution to improve accuracy.

II. BACKGROUND THEORY

A. RESERCH TREND

It is known that the synthesis of FCM and T2FS gives more room to handle the uncertainties of clustering caused by noisy environment. These hybrid algorithms include the general type-2 FCM [8], Interval Type-2 FCM (IT2-FCM) [9], kernel IT2-FCM [10], interval type-2 fuzzy c-regression clustering [11], interval type-2 possibilistic c-means clustering [12], [7], interval type-2 relative entropy FCM [13], particle swarm optimization based IT2-FCM [14], interval-valued fuzzy set-based collaborative fuzzy clustering [15]. This T2FS based algorithms have been successfully applied to areas like image processing, time series prediction and others.

Interval Type-2 FCM (IT2-FCM): In fuzzy clustering algorithms like FCM, the fuzzifier value m plays an important role in determining the uncertainty of clustering. However, the value of m is usually hard to be decided upon in advance. IT2-FCM considers the fuzzifier value as an interval $[m_1, m_2]$, and solves two optimization problems [16].

Another Type-2 fuzzy clustering Algorithm: Unlike IT2-FCM generates type-2 memberships by solving two optimization problems with two fuzzifier value, the another kind of Type-2 FCM (T2-FCM), whose type-2 membership is directly generated by extending a scalar membership degree to a T1FS. When restricting the secondary fuzzy sets to have triangular membership functions, T2-FCM extends a scalar membership u_{ij} to a triangular secondary membership function [17], [18].

B. FUZZIFIER VALUE

When the density or volume of each given cluster is different, the fuzzifier value plays a decisive role in finding the clustering membership function. It is assumed that the relative distances to the cluster center are all equal to 0.5, which means that the m of fuzzifier value is "1" and is considered a decision boundary. There is no fuzzy area under the above conditions.

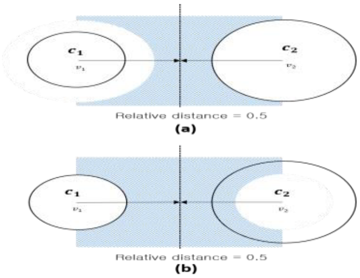

FIGURE 1. Fuzzy area between clusters according to m.

Figure 1(a) shows the case where a small m value is set in two clusters with different volumes. Since the section with a fuzzy membership value extends to a bulky C_2 cluster, applying it to the C_1 cluster allocates a lot of relatively unnecessary patterns. When a large m value is set as shown in (b) of Figure 1, it seems to have good performance because similar membership values are assigned, but the center value of the C_1 cluster tends to move to the C_2 cluster.

Assuming the points located at the centers v1 and v2 of the two clusters c1, and c2 and vertical lines.

Therefore, the membership function is calculated differently according to m and the membership value where the pattern belongs maybe different according to the membership function. Finally, when two fuzzifier value m_1, m_2 are used in interval type-2 fuzzy set (IT2 FS), the pattern can be classified more accurately than one fuzzifier value.

TABLE 1. Symbol for clustering method.

Symbol	Explanation
C	Bulky cluster
v	Cluster center
v_i, V	Cluster prototype
m	Fuzzifier value
u	Membership function
U	Partition matric
d_{ik}/d_{ij}	Euclidean distance value
δ	Threshold of fuzzify constant
$\tilde{A}$	Secondary membership degree
J	PFCM objective function
x	Input pattern
t_{ik}	Represents typicality,the input pattern k belongs to cluster i
γ_i	Scale defining point where typicality of the i-th cluster is 0.5
X_i	Input space
$\Phi(X_j)$	Kernel property space
K	Input space for kernel
S	Number of kernels
$\tilde{k}$	Gaussian multiple kernels
w_{il}	Resolution-specific weight
ρ	Gradient descent method

C. DETERMINING THE RANGE OF FUZZIFIER VALUE

As stated above, several methods have been proposed to determine the lowest and highest boundary range values of the fuzzifier value from given data [19]. The PFCM membership function for determining the range of the fuzzifier value is given as follows. The membership function at k-th data point for cluster i is shown in equation (1). d_{ik}/d_{ij} means Euclidean distance value between cluster and data point.

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} (d_{ik}/d_{ij})^{2/(m-1)}} \tag{1}$$

To determine the range of the fuzzifier value, the neighbor membership values are calculated, using the membership value obtained in (1). Summarization with an expression related to m is as equation (2) and the lowest and highest boundary values of the fuzzy constant can be obtained as C is the number of clusters and m is the fuzzifier value.

$$1 + \frac{C-1}{C} \cdot \frac{2}{\delta} \cdot |\Delta| \leq m \leq \frac{2\log d}{\log(\frac{\delta}{1-\delta} \cdot \frac{1}{c-1})} + 1$$
$$\text{where } \Delta = \frac{d_i - d_i^*}{d_i^*} \text{ and } \delta \text{ isthreshold.} \tag{2}$$

D. INTERVAL TYPE-2 FUZZY MEMBERSHIP FUNCTION

In general, the type-1 fuzzy set (T1 FS) has been widely used to represent pattern uncertainty in the field of pattern recognition. However, as previously shown, T1 FS cannot produce good result and be extended to type-2 fuzzy set (T2 FS) in order to control the uncertain fuzzifier value more efficiently [20], [21]. T2 FS, $\tilde{A}$, is represented as follows.

$$\tilde{A} = \int_{x \in X} u_{\tilde{A}}(x)/x = \int_{x \in X} \left[\int_{x \in j_x} f_x(u)/u \right] /x. \tag{3}$$

Expansion to T2 FS, which gives more control over uncertainty, generally yields better results than T1 FS. However, the calculation is complicated and requires a lot of computation [20]. To supplement excessive computation, IT2 FS with a secondary membership degree; 1 is used. IT2 FS, $\tilde{A}$,is expressed as equation (4). As seen from the equation, when the secondary membership degree is same at every point, it can be used as T1 FS.

$$\tilde{A} = \left[\int_{u \in jx} 1/u \right] /x \tag{4}$$

III. INTERVAL TYPE-2 PFCM ALGORITHM

IT2 PFCM is expressed as the sum of the weights of FCM method and PCM method and has both of above characteristics. Therefore, it is clustered in the direction of minimizing PFCM objective function as follow.

$$J_{m \cdot n}(U, T, V : X) = \sum_{k=1}^{n} \sum_{i=1}^{c} (a u_{ik}^{m} + b t_{ik}^{\eta}) \times \|x_k - v_i\|^2 + \sum_{i=1}^{c} \gamma_i \sum_{k=1}^{n} ((1 - t_{ik})^{\eta}) \tag{5}$$

$$\sum_{i=1}^{c} u_{ik} = 1, 0 < u_{ik}, t_{ik} \leq 1, m => 1, \eta > 1, \gamma > 0 \tag{6}$$

In equation (5), u_{ik} represents a membership value where the input pattern k belongs to cluster $\underline{i}$. x_i is the k-th input pattern, and v_i is the center value of the i-th cluster. m is a constant representing the degree of fuzziness and satisfies the condition of $m \in (1, \infty)$. t_{ik} represents typicality that the input pattern k belongs to cluster i, which is a feature of PFCM method using absolute distance. γ_i is a scale defining point where typicality of the i-th cluster is 0.5 and the value is an arbitrary number. To cluster with PFCM method, the objective function in above equation (5) should be minimized with respect to the membership function u_{ik}. Membership for this is obtained by equation (1). a and b are variables which determine the weight of FCM and PCM. In order to expand to IT2 FS, the uncertainty of the fuzzifier value m must be expressed. To draw m, you must create the lowest and highest membership functions using the primary membership function. The lowest and highest membership functions of PFCM according to m are as follows.

$$\bar{u}_{ik} = \left(\frac{1}{\sum_{j=1}^{1} \left(\frac{d_{ik}}{d_{ij}}\right)^{\frac{2}{m_1 - 1}}}\right) \tag{7}$$

$$\underline{u}_{ik} = \left(\frac{1}{\sum_{j=1}^{1} \left(\frac{d_{ik}}{d_{ij}}\right)^{\frac{2}{m_2 - 1}}}\right) \tag{8}$$

where m_1 and m_2 are the highest and lowest fuzzifier value, as shown in equation (5) representing objective function, the value γ_i also changes according to the lowest and highest membership functions. Using γ_i, the lowest and highest typicality is,

$$\bar{t}_{ik} = \begin{cases} \dfrac{1}{1+\left(\frac{d_{ik}^2}{\bar{\gamma}_i}\right)}, & if \dfrac{1}{1+\left(\frac{d_{ik}^2}{\bar{\gamma}_i}\right)^{\frac{2}{m_1-1}}} > \dfrac{1}{1+\left(\frac{d_{ik}^2}{\gamma_i}\right)^{\frac{2}{m2-1}}} \\ \dfrac{1}{1+\left(\frac{d_{ik}^2}{\gamma_i}\right)^{\frac{2}{m_2-1}}}, & otherwise \end{cases} \tag{9}$$

$$t_{ik} = \begin{cases} \dfrac{1}{1+\left(\frac{d_{ik}^2}{\bar{\gamma}_i}\right)}, & if \dfrac{1}{1+\left(\frac{d_{ik}^2}{\bar{\gamma}_i}\right)^{\frac{2}{m_1-1}}} \leq \dfrac{1}{1+\left(\frac{d_{ik}^2}{\gamma_i}\right)^{\frac{2}{m2-1}}} \\ \dfrac{1}{1+\left(\frac{d_{ik}^2}{\gamma_i}\right)^{\frac{2}{m_2-1}}}, & otherwise \end{cases} \tag{10}$$

After obtaining membership as above, the central value of each cluster must be updated. To update the center value, the type reduction process of changing type-2 fuzzy set to type-1 using the KM algorithm is performed and the updated center value of each cluster is as shown in equation (11).

$$v_i = \frac{\sum_{k=1}^{n} (au_{ik}^m + bt_{ik}^\eta) X_k}{\sum_{k=1}^{n} (au_{ik}^m + bt_{ik}^\eta)} \tag{11}$$

A. MULTIPLE KERNELS PFCM ALGORITHM

In general, the kernel method is to convert the input data from the input property space to the kernel property space through the kernel function using a space conversion function [21]. This is to change the kernel property space into the kernel property space making it easier to distinguish data that has or overlaps a non-linear boundary surface of input property space through kernel property space conversion.

If the data in the input space is X_i, $i = 1, \ldots, N$, the data converted to the kernel property space through the function is represented by $\Phi\left(X_j\right)$, $\mathrm{j} = 1 \ldots \mathrm{N}$.

Alike as general PFCM, in the case of Kernels-PFCM, the goal is to minimize the following objective function.

$$J^\phi = \sum_{k=1}^{n} \sum_{i=1}^{c} (au_{ik}^m + bt_{ik}^\eta) \times d_{ij}^2 + \sum_{i=1}^{c} \gamma \sum_{k=1}^{n} (1 - t_{ik})^\eta \tag{12}$$

In the input space for kernel K, the pattern x_i and the distance d_{ij} in the kernel attribute space of cluster prototype v_j are expressed as equation (13) by the kernel function.

$$\begin{aligned} d_{ij} &= \left\|\Phi(x_j) - \Phi(v_j)\right\|^2 \\ &= \Phi(x_j)\Phi(x_j) + \Phi(v_j)\Phi(v_j) - 2\Phi(x_j)\Phi(v_j) \\ &= K(x_j, x_j) + K(v_j, v_j) - 2k(x_j, v_j) \end{aligned} \tag{13}$$

In general, multiple kernels with the number of kernels S are assumed so new Gaussian multiple kernels $\tilde{k}$ using Gaussian kernel are as follows [21].

$$\tilde{k}^{(j)} = (x_j, v_j) = \sum_{l=1}^{s} \frac{w_{il}}{\sigma_l} \frac{\exp\left(-\frac{\|x_j - v_j\|^2}{2\sigma_l^2}\right)}{\sum_{t=1}^{s} \frac{w}{\sigma_t}} \tag{14}$$

From [22] method, normalized kernel is defined so that e FCM-MK is to identify the resolution-specific weight, the membership values and the cluster prototypes. Using this optimization method, following PFCM objective function should be minimized. Resolution-specific weight w_{il}, membership value u_{ij} and cluster Prototype v_i are determined by minimizing the objective function below.

$$\begin{aligned} &J_{m,\eta}(U, T, V; X) \\ &\quad = 2 \sum_{k=1}^{n} \sum_{i=1}^{c} (au_{ik}^m + bt_{ik}^\eta \\ &\qquad \times \left(1 - \sum_{l=1}^{s} \frac{w_{il}}{\sigma^2} \exp\left(-\frac{\left\|x_j - v_i\right\|^2}{2\sigma_l^2}\right) \times \frac{1}{\sum_{t=1}^{s} \frac{w}{\sigma_t}}\right) \\ &\qquad + \sum_{i=1}^{c} \gamma_i \sum_{k=1}^{n} ((1 - t_{ik})^\eta) \end{aligned} \tag{15}$$

B. MULTIPLE KERNELS INTERVAL TYPE-2 PFCM ALGORITHM

In order to solve the uncertainty existing in the fuzzifier value m in the general PFCM algorithm, Multiple Kernels PFCM algorithm should be extended to the Interval Type-2 fuzzy set. If there are N data, C clusters, U partition matric, V set of cluster prototype, W set pf resolution-specific weight and S kernels, the cluster prototype can be obtained by minimizing the Gaussian kernel objective function as follows.

$$J(U, V, W) = 2 \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^{m} d_{ij}^{2} \tag{16}$$

where,

$$d_{ij}^{2} = \left(2 - 2 \sum_{i=1}^{S} \frac{w_{il}}{\sigma_l} \frac{\exp\left(- \frac{\|x_j - v_j\|^2}{2\sigma_l^2} \right)}{\sum_{t=1}^{s} \frac{w}{\sigma_t}} \right) \tag{17}$$

The cluster prototype is calculated by optimizing the objective function for the center v_i of the cluster [22].

$$v_i = \left(2 - 2 \sum_{i=1}^{S} \frac{w_{il}}{\sigma_l} \frac{\exp\left(- \frac{\|x_j - v_j\|^2}{2\sigma_l^2} \right)}{\sum_{t=1}^{s} \frac{w}{\sigma_t}} \right) \tag{18}$$

where

$$\bar{K}^{(i)}(x_j, v_i) = \left(\sum_{i=1}^{S} \frac{w_{jl}}{\sigma_l^3} \frac{\exp\left(\|x_j - v\|^2 \right)}{\sum_{t=1}^{S} \frac{w}{\sigma_t}} \right) \tag{19}$$

Calculated the smallest membership value and the largest membership value for each pattern using the Interval Type-2 fuzzy set, optimized membership value, is used for calculating the crisp value v_i. To calculate v_R and v_L, it is necessary to determine the upper or lower bound of membership. It is organized as follows by given formula [23].

For v_L,

if $(v(i < k))$ then $u_{ij} = \bar{u}_{ij}$

else $\boldsymbol{u_{ij}} = _{\boldsymbol{ij}}$

$$v_{iL} = \frac{\sum_{j=1}^{N} u_{ij}^{m} \bar{K}^{(i)}(x_j, v_i) x_j}{\sum_{j=1}^{N} u_{ij}^{m} \bar{K}^{(i)}(x_j, v_i)} \tag{20}$$

For v_R,

if $(v(i < k))$ then $\boldsymbol{u_{ij}} = \bar{u}_{\boldsymbol{ij}}$

else $\boldsymbol{u_{ij}} = _{\boldsymbol{ij}}$

$$v_{iR} = \frac{\sum_{j=1}^{N} u_{ij}^{m} \bar{K}^{(i)}(x_j, v_i) x_j}{\sum_{j=1}^{N} u_{ij}^{m} \bar{K}^{(i)}(x_j, v_i)} \tag{21}$$

Using the final v_R and v_L, the crisp center value is obtained from defuzzification as follows.

$$v_i = \frac{v_{iL} + v_{iR}}{2} \tag{22}$$

Using the cluster Prototype v_i, obtained through the optimization function and the membership value u_{ij}, the resolution-specific weight value w_{il} is updated as follows.

$$w_{il}^{(new)} = w_{il}^{(old)} - \rho \frac{\partial J}{\partial w_{il}} \tag{23}$$

where

$$\frac{\partial J}{\partial w_{il}} = -2 \sum_{i=1}^{N} \frac{u_{ij}^{m}}{\sum_{t=1}^{S} \frac{w_t}{\sigma_t}} \left(K(x_j, v_i - \bar{K}^{(i)}(x_i, v_j) \right) \tag{24}$$

Here, ρ is a gradient descent method as learning rate parameter. Finally, clustering is performed through type-reduction and hard patitioning as described in Interval Type-2 PFCM [24].

IV. DETERMINATION OF FUZZIFIER VALUE USING HISTOGRAM

The proposed method in this paper extracts information from data given through the histogram method, and then adaptively calculates the fuzzifier value based on the obtained information. First, the IT2 PFCM algorithm, defined in the previous section, estimates roughly which cluster the data belongs to and then obtains a histogram based on the data from the classified clusters. The histogram obtained in IT2 PFCM is made into a gentler and smoother histogram through the triangular window and the membership function is obtained by using a curve fitting on top of this histogram. To get the IT2 FS, you need to determine the FOU, which is the set of all major memberships of the T2 FS. Therefore, the values of the histogram greater than the membership value are assigned as the histogram of the highest membership and the values of the histogram with values less than the membership value are stored as the histogram of the lowest membership. The lowest and highest membership functions can be obtained again using the curve-fitted histogram. Curve fitting is implemented separately on upper and lower histograms giving us upper and lower membership values. We propose an algorithm that estimates the fuzzifier value m_1, m_2 using the membership function. Figure 2 shows histograms and FOU examples determined by class and dimension.

To find X that satisfies function $f(X) = 0$, it can be expressed in the form of $X = g(X)$ using fixed-point iteration, and the following X is

$$X_{i+1} = g(X), \quad i = 0, 1, \ldots, N \tag{25}$$

Equations (7) and (8) of the membership function u_i are expressed in the form of equation (25) as follows.

$$u_i = \frac{1}{\sum \left(\frac{d_{ik}}{d_{ij}} \right)^{\frac{2}{m-1}}} \tag{26}$$

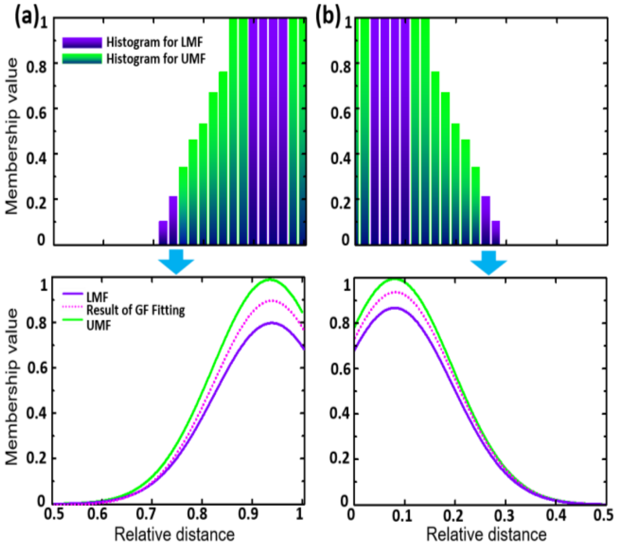

FIGURE 2. FOU obtained for individual class and dimension class 1 dimension 1, and (b) class 2 dimension 1.

If you take the log on both sides in equation (26), equation (27) can be summarized as follows:

$$\log\left(\frac{1}{u_1}\right) = \frac{2}{m-1}\log\left(\frac{d_{ki}}{d_{1i}}\right) + \log\left(1+\sum_{j=2}^{c}\left(\frac{d_{ki}}{d_{ji}}\right)^{\frac{2}{m_{dd}-1}}\right). \quad (27)$$

Rearranging (27) and expressing it in terms of m, gives us (28), (29).

$$\gamma = \frac{\log\left(\frac{1}{u_j}\right) - \log\left(1+\sum_{k=2}^{C}\left(\frac{d_{ij}}{d_{ik}}\right)^{2/m_{old}-1}\right)}{\log\left(\frac{d_{ij}}{d_{ik}}\right)} \quad (28)$$

$$m_{jnew} = 1 + \frac{2}{\gamma} \quad (29)$$

As in the above process, the value of $u_i \in \{\underline{u_i}(X_k), \overline{u_i}(X_k)\}$ and m_{new} is used as a function to get the membership value. When equation (10) is applied to all cluster data and calculated, new m_{1inew} and m_{2inew} values can be obtained [25]. By taking the average value of the fuzzifier value obtained through equation (29), the new fuzzifier value m_1 and m_2 are finally determined, and the new fuzzifier value obtained for clustering are as follows.

$$m_1 = \left(\sum_{i=1}^{N} m_{1i}\right)/N, \quad m_2 = \left(\sum_{i=1}^{N} m_{2i}\right)/N \quad (30)$$

Figure 2 shows that histograms and FOU examples are determined by class and dimension. Upper membership function (UMF) histogram and lower membership function (LMF) histogram are obtained according to class and dimension. A new membership function from the Gaussian Curve Fitting (GF-F) method can be applied to calculate the adaptive fuzzifier value.

V. APPLICATION TO SENSOR DATA

In order to check the performance of the proposed algorithm, supervised learning and unsupervised learning were compared for sensor data of various characteristics. SVM method for supervised learning, K-means for unsupervised learning and Interval Type-2 PFCM (IT2PFCM), the proposed algorithm was tested and compared. Supervised learning is different from the proposed algorithm but presented as the need to apply a context-sensitive test method. m_1 and m_2 values were tested in the range of 1 to 5. The σ value was fixed to the most common Gaussian function value.

Acquired sensor data measured by indoor temperature / humidity, VOCs, and miscellaneous dust (PM 10, 2.5) sensors. This data is used as training data as labeling data composed of outlier and spatial information, as well as continuous data that has not been labeled. The volume of sensor training and validating data is about 600,000 cases with various sensor data for 2 weeks, and test data used for prediction is about 150,000 cases, consist of 80%: 20%.

It is performed by selecting the number of classes (K) to be clustered and determining one cluster centroid per sensor in principle. However, cluster expansion is possible according to various event conditions. In this case, you can find outliers out of the centroid (center point) and find the ratio by the number of outliers. Finally, the accuracy of clustering is calculated by which cluster the input data of the proposed algorithm belongs to. It is useful to use this algorithm when expanding to a smart city. This is because numerous sensors and sensor big data cannot be managed individually, and numerous data received in real time cannot be labeled individually.

Real-time data acquisition from multiple sensor devices is displayed on a chart, and the proposed algorithm determines the located cluster of each data, eventually, present the accuracy of clustering as a result. The training and validation data is classified to 3 features, and each one consists of 10 or more pattern data. It was clustered using the features of this sensor data.

Instance of the fuzzy area according to the value of m_1 and m_2 using the characteristics of the Interval type-2 membership set, uncertainty can be reduced and an appropriate fuzzy area for the cluster volume can be formed, it is as shown in the figure 3.

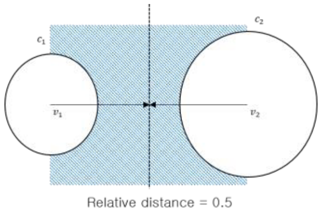

FIGURE 3. Instance of appropriate fuzzy area using Interval type-2.

To expand to the Interval Type-2 fuzzy set and express uncertainty for m, the input pattern, which is the primary fuzzy set, is assigned to the Interval Type-2 fuzzy set. To this end, the upper membership function and the lower

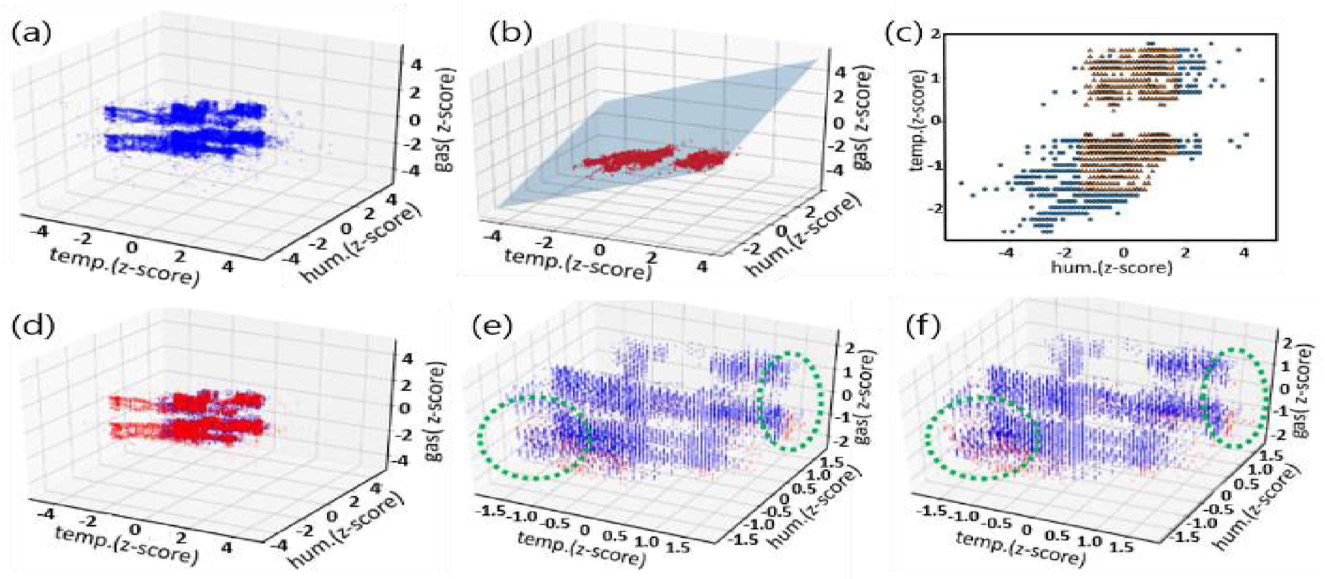

FIGURE 4. **Coordinates of raw sensor data and supervised learning, IT2 PFCM, the proposed algorithm.**

membership function are created as primary membership functions.

After obtaining the upper and lower membership for each cluster, we need to update the center values for each cluster. In this process, the membership is a Type-2 fuzzy set, however the center value is a crisp value, the value cannot be obtained using the above method. Therefore, in order to update the center value, Type reduction is performed by changing the Type-2 fuzzy set to the Type-1 fuzzy set. In addition, defuzzification should be accomplished to change the value of Type-1 to a crisp value.

The Upper Membership Function (UMF) Histogram and Lower Membership Function (LMF) Histogram are drawn. A new membership function defined from the Gaussian Curve Fitting (GF-F) method is obtained. Footprint of Uncertainty (FOU), set of major memberships, is determined, finally, new fuzzifier value m_1 and m_2 are derived. m is a value that determines the degree of final clustering fuzzifier as the value of the fuzzy parameter.

When comparing and testing several existing AI algorithms with the proposed algorithm, (a) in Figure 4 is the raw data displaying the sensor data on the coordinate plane. All presented data were converted to z-score. z-score is a value that statistically creates a normal distribution and shows where each data is located on the standard deviation. The standard value indicates how far away the data is from the mean, negative; below the average, positive; above the average.

(b) is processed by SVM method in supervised learning and (c) is shown in cross section. SVM is one of the classification algorithms, and is a good classification rate algorithm. Among the two types of linear and nonlinear algorithms, this analysis classifies sensor data using a linear algorithm. In order to improve readability in 3D coordinates reflecting 3 characteristics, it was presented as a cross-section.

(d) is a comparison between raw data(red) and outlier removed data (blue). As the outliers are removed, the original data is relatively centered. The outlier removes 5.275% of the total data with a standard score of 2.0 or higher. For clustering accuracy, K-means performed better than PFCM. There is no significant difference in performance, so picture representation is excluded.

(e) is an enlarged comparison of the outlier removed data (blue) and the result processed by the IT2 PFCM algorithm (red). When compared with the data with outliers removed, the IT2PFCM algorithm centralizes the sensor data into clusters.

Finally, (f) is an enlarged comparison of the outlier removed data (blue) and the result of processing with the proposed algorithm (red). In the coordinate plane, (e) and (f) do not seem to differ significantly, but it can be seen from the numerical values that the proposed algorithm is improved over the IT2 PFCM.

In Figure 4 (f), the values of m_1 and m_2 are adaptively found through the proposed algorithm that finds the most suitable fuzzifier value by conventional learning. the m_1 and m_2 applied to the proposed algorithm to get new clusters and the above process is repeated iteratively till the termination criteria are satisfied. If there is negligible change in the resulting fuzzifier values or termination criteria is satisfied then end the algorithm, otherwise, the algorithm is again repeated from stepwise approach using these new m_1 and m_2. The stepwise approach is setting initial values, applying to m_1 and m_2 in the proposed algorithm, calculating membership of each data point, generating a histogram and curve fitting according to this value of membership.

TABLE 2. Comparison with proposed algorithms, such as accuracy.

	Supervised Learning (SVM)	Unsupervised Learning (K-means)	PFCM	IT2 PFCM (m=1,2)	**The Proposed Algorithm (m=2.2, 2.8)**
Total Accuracy of clustering (%)	78.5	79.10	78.03	81.02	**85.53**
Temperature Accuracy (RMSE)	70.1	82.55	81.23	83.67	**87.58**
Humidity Accuracy (RMSE)	83.8	80.47	78.05	81.87	**86.21**
Gas Accuracy (RMSE)	42.3	74.27	74.80	77.51	**82.76**
Hyper parameter (Iteration) (Batch size)	1,000	100 (247451)	100 (245400)	100 (239301)	**100 (233228)**
(Input neurons : 30), (Output neurons : 1), (Hidden neurons : 15), (Epoch : 13), (Hidden layers : 11)					

In order to perform the test in a consistent situation, some hyper parameters were set identically. As shown in Table 2, it shows the clustering accuracy for temperature / humidity / gas / total sensor data, and was improved in the order of SVM [26] – PFCM [27] – K-means [28] – IT2PFCM – The proposed algorithm. The accuracy of the gas data was predicted to be lower than that of the other data, which is because it reflects the characteristics of gas sensors based semiconductor that have various environmental effects. Finally, it was tested that the proposed algorithm performance improved over the IT2PFCM algorithm.

VI. CONCLUSION

Dealing complex data with noise, the membership value is subdivided into the upper / lower membership value and introduced into histogram and Gaussian Kernel method to improve the accuracy of the Interval type-2 Possibilistic Fuzzy C-means Multiple. It plays an important role in deriving the better fuzzifier value *m*. This theoretical proof is verified by practical data. In practical situation, IT2 PFCM with new method compared to existing algorithms shows 95.6 $\sim$ 5.6% improvement in accuracy. In further study, deriving parameters such as various weight values using the above method should be carried out to stabilize accuracy of clustering and improve performance.

REFERENCES

[1] G. Raju, B. Thomas, S. Tobgay, and S. Kumar, "Fuzzy clustering methods in data mining: A comparative case analysis," in *Proc. Int. Conf. Adv. Comput. Theory Eng.*, Dec. 2008, pp. 489–493.

[2] H. Yu and J. Fan, "Cutset-type possibilistic C-means clustering algorithm," *Appl. Soft Comput.*, vol. 64, pp. 401–422, Mar. 2018.

[3] K. D. Koutroumbas, S. D. Xenaki, and A. A. Rontogiannis, "On the convergence of the sparse possibilistic C-means algorithm," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 1, pp. 324–337, Feb. 2018.

[4] M. R. N. Kalhori and M. H. F. Zarandi, "Interval type-2 credibilistic clustering for pattern recognition," *Pattern Recognit.*, vol. 48, no. 11, pp. 3652–3672, Nov. 2015.

[5] J. Zhou, Z. Lai, C. Gao, D. Miao, and X. Yue, "Rough possibilistic C-means clustering based on multigranulation approximation regions and shadowed sets," *Knowl.-Based Syst.*, vol. 160, pp. 144–166, Nov. 2018.

[6] T. Wang and W.-L. Hung, "A generalized possibilistic approach to shell clustering of template-based shapes," *J. Stat. Comput. Simul.*, vol. 87, no. 3, pp. 423–436, Feb. 2017.

[7] E. Rubio, O. Castillo, F. Valdez, P. Melin, C. I. Gonzalez, and G. Martinez, "An extension of the fuzzy possibilistic clustering algorithm using type-2 fuzzy logic techniques," *Adv. Fuzzy Syst.*, vol. 2017, pp. 1–23, Jan. 2017, doi: 10.1155/2017/7094046.

[8] O. Linda and M. Manic, "General type-2 fuzzy C-means algorithm for uncertain fuzzy clustering," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 5, pp. 883–897, Oct. 2012.

[9] C. Hwang and F. C.-H. Rhee, "Uncertain fuzzy clustering: Interval type-2 fuzzy approach to *C*-means," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 1, pp. 107–120, Feb. 2007.

[10] P. Kaur, I. M. S. Lamba, and A. Gosain, "Kernelized type-2 fuzzy C-means clustering algorithm in segmentation of noisy medical images," in *Proc. IEEE Recent Adv. Intell. Comput. Syst.*, Sep. 2011, pp. 493–498.

[11] M. H. F. Zarandi, R. Gamasaee, and I. B. Turksen, "A type-2 fuzzy C-regression clustering algorithm for Takagi–Sugeno system identification and its application in the steel industry," *Inf. Sci.*, vol. 187, pp. 179–203, Mar. 2012.

[12] M. A. Raza and F. C.-H. Rhee, "Interval type-2 approach to kernel possibilistic C-means clustering," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Jun. 2012, pp. 1–7.

[13] M. Zarinbal, M. H. F. Zarandi, and I. B. Turksen, "Interval type-2 relative entropy fuzzy C-means clustering," *Inf. Sci.*, vol. 272, pp. 49–72, Jul. 2014.

[14] E. Rubio and O. Castillo, "Interval type-2 fuzzy possibilistic C-means optimization using particle swarm optimization," in *Nature-Inspired Design of Hybrid Intelligent Systems*, vol. 667. Springer, 2017, pp. 63–78.

[15] L. T. Ngo, T. H. Dang, and W. Pedrycz, "Towards interval-valued fuzzy set-based collaborative fuzzy clustering algorithms," *Pattern Recognit.*, vol. 81, pp. 404–416, Sep. 2018.

[16] X. Yang, F. Yu, and W. Pedrycz, "Typical characteristics-based type-2 fuzzy C-means algorithm," *IEEE Trans. Fuzzy Syst.*, early access, Jan. 28, 2020, doi: 10.1109/TFUZZ.2020.2969907.

[17] J. P. Sarkar, I. Saha, and U. Maulik, "Rough possibilistic type-2 fuzzy C-means clustering for MR brain image segmentation," *Appl. Soft Comput.*, vol. 46, pp. 527–536, Sep. 2016.

[18] L. Maciel, R. Ballini, and F. Gomide, "Evolving possibilistic fuzzy modelling," *J. Stat. Comput. Simul.*, vol. 87, no. 7, pp. 1446–1466, May 2017.

[19] I. Ozkan and I. Turksen, "Upper and lower values for the level of fuzziness in FCM," in *Fuzzy Logic*. Springer, 2007, pp. 99–112.

[20] F. Chung and H. Rhee, "Uncertain fuzzy clustering: Insights and recommendations," *IEEE Comput. Intell. Mag.*, vol. 2, no. 1, pp. 44–56, Feb. 2007.

[21] S. Majeed, A. Gupta, D. Raj, and F. C.-H. Rhee, "Uncertain fuzzy self-organization based clustering: Interval type-2 fuzzy approach to adaptive resonance theory," *Inf. Sci.*, vol. 424, pp. 69–90, Jan. 2018.

[22] N. Baili and H. Frigui, "Fuzzy clustering with multiple kernels," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jun. 2011, pp. 490–496.

[23] D. S. Comas, G. J. Meschino, A. Nowé, and V. L. Ballarin, "Discovering knowledge from data clustering using automatically-defined interval type-2 fuzzy predicates," *Expert Syst. Appl.*, vol. 68, pp. 136–150, Feb. 2017.

[24] Abhishek, A. Jeph, and F. C.-H. Rhee, "Interval type-2 fuzzy C-means using multiple kernels," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2013, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/6622306

[25] W.-H. Joo and F. C.-H. Rhee, "Determining the fuzzifier values for interval type-2 possibilistic fuzzy C-means clustering," *J. Korean Inst. Intell. Syst.*, vol. 27, no. 2, pp. 99–105, Apr. 2017.

[26] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Netw.*, vol. 17, no. 1, pp. 113–126, Jan. 2004.

[27] V. Kalist, P. Ganesan, B. S. Sathish, J. M. M. Jenitha, and K. B. Shaik, ''Possiblistic-fuzzy C-means clustering approach for the segmentation of satellite images in HSL color space,'' *Procedia Comput. Sci.*, vol. 57, pp. 49–56, Jan. 2015.

[28] A. K. Jain, ''Data clustering: 50 years beyond K-means,'' *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

JAEHYUK CHO received the Ph.D. degree in computer science (mobile and embedded computing systems) from Chung-Ang University, South Korea.

He is currently working as a Professor with the Department of Electronic Engineering, Soongsil University. He is also a National Research and Development Program Project Manager with the Korea Institute of S&T Evaluation and Planning (KISTEP). His research interests include AI, data process, bigdata of sensors, the IoT, smart city, SW platform systems, and embedded systems. He has served as an ICONIC Board Member. He has also served as a Reviewer and PlatCon.

WONHEE JOO received the Ph.D. degree in electronic and communication engineering from Hanyang University, South Korea.