Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 17, No. 3, 2026

Strict Standard Normal Cloud Sets C-Means Clustering: A New Approach for Competitive Intelligence Classification of Hydrogen Energy

Hongli Wang *

Academic Affairs Office, Fujian Jiangxia University, Fuzhou, Fujian, 350000, China
*** Corresponding author Email:** graduated852@163.com

Abstract: This research first defines strict standard normal cloud sets—a mathematical model integrating randomness and fuzziness to quantify and address dataset uncertainty—and develops the corresponding cloud generation algorithm. This model converts quantitative feature data into cloud sets (characterized by parameters Ex, En, He, β), effectively capturing the uncertain distribution of features. In the classification of hydrogen energy technology competitive intelligence, this model is applied: first extract key features, then use the proposed model to generate cloud sets for each feature. We further derive the iterative formulas of strict standard normal cloud sets C-means clustering, including formulas for updating cluster center cloud parameters and calculating membership degrees. Additionally, we use the feature cloud sets to construct the initial membership degree matrix, and select initial clustering centers based on the membership degree values. Comparative experiments show that this method outperforms traditional fuzzy C-means in handling text data related to intelligence, with higher classification accuracy. It provides a rigorous technical framework for the precise classification of hydrogen energy technology competitive intelligence, supporting decision-making in industrial dynamics analysis.

Keywords: Competitive Intelligence Classification; C-Means Clustering; Hydrogen Energy Technology; Strict Standard Normal Cloud Sets; Randomness and Fuzziness.

1. Introduction

K-means employs hard clustering, where each data point belongs to exactly one cluster. In contrast, C-means adopts soft clustering, enabling each data point to have respective membership degrees corresponding to multiple clusters. Chakraborti and Dey developed the multi-level *K*-means text clustering technique in 2016, and illustrated its application in text clustering for competitive intelligence through systematic examples, providing a reliable framework for related research directions [1]. In existing research on text clustering for competitive intelligence, the uncertain C-means clustering method has not yet been widely adopted. Existing studies on uncertain C-means clustering have proposed several approaches, including fuzzy C-means clustering, grey C-means clustering, and cloud model-based C-means clustering, etc.

Focusing specifically on fuzzy C-means (FCM) clustering, existing studies have explored a range of theoretical frameworks and practical implementations. In 1984, Bezdek and colleagues published a seminal paper that further standardized the FCM algorithm, accompanied by detailed implementation procedures[2]. This work established a clear framework for the algorithm's convergence analysis and parameter conFigureuration, thereby facilitating its widespread adoption in engineering applications. At its core, FCM aims to minimize an objective function, with the final clustering results derived through iterative optimization. In 1993, Krishnapuram and Keller introduced possibilistic C-means (PCM) clustering, which integrates possibility theory with fuzziness[3]. By relaxing the constraint that the sum of membership degrees equals one and incorporating the concept of typicality, this approach achieves a remarkable enhancement in robustness against noise and other centers. Specifically, the algorithm modifies the objective function to minimize the weighted distances from data points to cluster centroids while maximizing intra-cluster compactness, thus overcoming the inherent limitation of FCM in terms of noise sensitivity.

To enhance the uncertainty characterization capability of conventional FCM, Zeng and colleagues extended the algorithm by modeling membership degree, non-membership degree, and hesitation degree via Intuitionistic Fuzzy Sets (IFSs) in 2020, thereby proposing the Hesitant Fuzzy C-Means (HFCM) Algorithm and its application in image segmentation [5]. In 2018, Kim and co-workers integrated Interval Type-2 Fuzzy Sets (IT2FSs) into the FCM framework to develop a fuzzy classifier, where the inherent uncertainty in data is effectively addressed through the employment of interval-valued membership degrees[6]. To address the sensitivity of conventional IT2FCM to the fuzzifier parameter, Wu and Guo proposed a robust single-fuzzifier interval type-2 fuzzy local C-means clustering approach in 2021, which leverages adaptive interval-valued data for land cover segmentation[6]. In the same year, Xu and coauthors proposed the Reliability-based Spatial Context Fuzzy C-Means (RSCFCM) algorithm, which compensates for the inherent limitation of existing FCM variants—their susceptibility to noise—when applied to image segmentation[7].

Despite the incremental improvements achieved by the aforementioned methodologies in this direction, several notable limitations still merit focused attention for further refinement. Firstly, the membership degree in FCM is characterized by "Type-1 fuzzy sets," where each sample's degree of belonging to a cluster is a definite value within the interval [0,1]. Its limitation lies primarily in the inherent trade-off between "determinism" and "flexibility." In practical

scenarios, however, data noise and measurement errors introduce uncertainty into the membership degrees themselves. Such "black-or-white" precise membership degrees fail to capture the "fuzziness within fuzziness" inherent in data ambiguity, thereby reducing adaptability to complex uncertain datasets. Secondly, in FCM, membership degrees are computed solely based on the Euclidean distance between samples and cluster centers, with closer distances corresponding to higher degrees. This singular association logic overlooks data characteristics such as local density and distribution patterns, leading to a disconnect between membership assignment and the true data relatedness in non-convex datasets or those with uneven density. Thirdly, conversely, the membership degree in IT2FCM is defined by "Type-2 fuzzy sets," where each sample's membership is itself a fuzzy interval. Its key limitation manifests as a conflict between "complexity" and "practicality." Currently, there exists no unified criterion to evaluate the "reasonableness of interval width": excessively narrow intervals reduce IT2FCM to a mere equivalent of FCM, while overly wide intervals blur membership information—ultimately undermining cluster discriminability. Fourthly, notably, both methods employ fixed rules for membership calculation, whether through distance-based association or fuzzy coefficient control, and lack the ability to adaptively adjust to data distribution characteristics (e.g., convex/non-convex, uniform/non-uniform density). Additionally, their membership computation considers only the distance between individual samples and cluster centers, neglecting the contextual relationships between samples and their neighboring data points.

About Grey C-means (GCM) clustering, existing investigations have primarily focused on addressing uncertainty-related issues in data processing under conditions of poor information. In 2016, Li and colleagues pioneered the integration of grey relational analysis into FCM, proposing the concept of "balanced proximity" to replace Euclidean distance [8]. In 2021, Liu and Zeng Perform automatic clustering on the samples in the historical database based on WGRA-FCM sample similarity measurement [9]. In 2025, Li proposed the "trend probability correlation degree," a novel index designed to resolve the inaccuracy of traditional grey correlation degrees in quantifying fluctuating sequences. By integrating this innovative index with an improved grey correlation matrix, a grey principal component clustering model was constructed. Empirical application in regional economic indicator clustering shows that this model outperforms conventional approaches in trend identification [10].

Despite its merits, the Grey C-Means (GCM) clustering algorithm has several notable limitations. Specifically, GCM merely replaces the Euclidean distance used in traditional C-means variants with grey relational degree for similarity calculation, without explicitly defining how "grey degree" quantifies its impact on the clustering process. For instance, the "degree of incomplete information" of data is not deeply integrated into the mechanisms of cluster center update or membership assignment. As a result, "greyness" serves only as a tool for similarity computation rather than a core mechanism for handling uncertainty, leading to indistinct essential differences from conventional clustering algorithms. During the algorithm's iterative process, membership degrees and cluster centers are adjusted solely based on grey relational degree, with no inverse correction of "grey degree evaluation" based on intermediate clustering results. For example, poor clustering performance of a sample group may stem from high greyness of the corresponding data, yet GCM fails to adjust the grey degree weight for this group accordingly. Subsequent iterations still rely on initial grey degree calculations, lacking the ability to dynamically adapt to changes in data uncertainty. Furthermore, grey theory targets uncertainty characterized by "partial information known and partial information unknown," yet GCM does not account for data noise, Other centers, or fuzziness. For instance, when processing noisy data, GCM does not implement special treatments for the greyness or relational degree of noisy samples, causing noise to directly interfere with cluster center computation and induce deviations in membership assignment—a drawback shared with FCM's inherent sensitivity to noise.

Regarding Cloud Model-based C-means (CCM) clustering, the current research landscape is characterized by diverse methodological improvements and application extensions. In 2007, Qin et al. proposed a cloud model-based fuzzy C-means clustering algorithm. As an effective transformation framework between quantitative data and qualitative concepts, this method was applied to the field of image segmentation, with experimental results showing that it can generate well-defined initial cluster centers [11]. In 2008, Yang et al. proposed a cloud model-based C-means clustering approach, where each cluster is modeled using a cloud model to capture both fuzziness and randomness inherent to the cluster. This approach enhances the adaptability of the clustering process compared to FCM and avoids the trivial solution of the objective function in FCM by eliminating the algorithm's inherent normalization constraint [12]. In 2014, Song et al. give the extraction method of power load patterns based on cloud model and fuzzy clustering [13]. In 2023, Liu et al. introduced "cloud-cluster," which incorporates the "cluster concept drift degree" to assess the uncertainty of concepts during the clustering process [14]. In 2024, Mao and Xu improved classification performance by adopting a strategy focused on mitigating uncertainty associated with boundary data through cloud models [15]. In 2025, Anjali and Gupta integrated Shannon entropy into cloud models to improve model accuracy and decision-making effectiveness, proposing a Shannon entropy-based backward cloud transformation algorithm [16]. These studies demonstrate the extensive applications of CCM, yet its inherent limitations in membership degree calculation and result reproducibility have not been adequately addressed, hindering its further adoption in complex practical scenarios.

The core limitation of Cloud Model-based C-means (CCM) clustering lies in its handling of membership degrees. CCM attempts to optimize clustering using the "randomness-fuzziness" dual characteristics of cloud models, yet the complex logic of cloud generators leads to uncontrolled membership degree calculation and increased parameter dependence. When forward cloud generators generate cloud droplets, membership degrees (certainty degrees of cloud droplets) are generated via normal distribution random functions. This means that even for the same dataset and identical parameters, the membership degrees generated in each run will differ due to inherent randomness—resulting in unstable clustering outcomes and compromised reproducibility. In contrast, membership degrees in conventional FCM are deterministic values, ensuring result replicability. Furthermore, the membership degree in CCM

simultaneously embodies "fuzziness" (the degree of a sample's affiliation to a cluster) and "randomness" (the probability distribution of cloud droplets), yet the algorithm fails to explicitly define the weight relationship between these two attributes. For example, a sample with a membership degree of 0.7 (certainty degree) provides no clarity on whether this value is dominated by "similarity to the cluster center" (fuzziness-driven) or "random cloud droplet distribution" (randomness-driven), rendering the physical meaning of membership degrees more ambiguous than that of FCM.

Regarding the inheritance and development of uncertainty modeling, the research context unfolds as follows. Zadeh was the first to propose fuzzy set theory, initially using a single membership degree to characterize uncertainty [17]. To address the limitation that "fuzzy sets cannot quantify the uncertainty inherent to membership degrees themselves," type-2 fuzzy sets were subsequently proposed [18–22] and have been widely applied in various fields, such as approximate reasoning [23], mobile robot obstacle avoidance [24], customer satisfaction sensory evaluation [25], decision-making [26], and image enhancement [27]. While type-2 fuzzy sets account for the uncertainty of membership degrees, they still rely on the mathematical framework of "fuzzy sets" and require predefined membership functions, without integrating "objective randomness." In 1995, Li Deyi proposed the cloud model, which integrates fuzziness and randomness to comprehensively and organically characterize uncertainty [28–30]. This model has been widely applied in diverse fields, including spatial data mining [31], web log data mining [32], decision-making methods [33], longitudinal control [34], and linguistic petri nets [35], etc. The cloud model addresses the critical limitations of type-2 fuzzy sets—specifically their separation of fuzziness from randomness and reliance on predefined functions—by constructing an "fuzzy-random integrated" uncertainty representation model through a "three-parameter cloud structure + random cloud droplet generation." However, the cloud model lacks a mathematical framework for "set relationships" (e.g., intersection, union, and complement operations), making it difficult to handle decision-making or reasoning scenarios involving multi-concept collaboration. In 2020, Wang proposed cloud set theory, which integrates cloud models into a set framework while retaining randomness [36]. In 2022, Wang presented the cloud relationship matrix and cloud relationship formula [37]. Subsequently, he demonstrated the applications of cloud sets in uncertain linear programming [38] and intelligent control decision-making [39], and proposed uncertain linear regression with cloud set coefficients [40]. Addressing the core limitation of cloud models—being only capable of handling single concepts and lacking set operation capabilities—cloud set theory upgrades cloud models from "single-concept representation tools" to "a mathematical framework for multi-concept collaborative processing" through "cloud set definition + relationship matrix + constraint optimization." This expansion broadens their applications in decision-making, optimization, and reasoning scenarios.

Based on the aforementioned literature, the evolution from fuzzy sets to type-2 fuzzy sets, cloud models, and further to cloud set theory revolves around the core goal of "more accurately characterizing uncertainty in the real world." This evolution is centered on two key advancements: progressively expanding the coverage of uncertainty (from fuzziness to the unification of fuzziness, randomness, and set relationships) and enhancing quantitative application capabilities (from qualitative characterization to optimized decision-making). The evolutionary pathway—fuzzy sets (single membership degree, characterizing fuzziness) → type-2 fuzzy sets (fuzzy sets of membership degrees, characterizing fuzziness of fuzziness) → cloud models (three parameters + cloud droplets, characterizing fuzzy-random unification) → cloud set theory (cloud set operations + relationship matrices, characterizing fuzzy-random-set relationship unification)—reflects a gradual deepening in the understanding and handling of uncertainty.

To summarize, existing uncertainty-based C-means clustering methods have the following limitations: (1) Traditional fuzzy C-means clustering methods rely on fixed fuzzy membership degrees, introducing significant subjectivity and failing to fully reflect the scope of real-world uncertainty. Type-2 fuzzy set-based C-means clustering methods address the uncertainty of membership degrees but still rely on predefined membership functions and lack integration with objective randomness. (2) Grey C-means clustering methods are suitable for handling clustering problems under conditions of poor information but fail to deeply integrate the "grey degree" with cluster center update and membership assignment mechanisms, resulting in indistinct differences from conventional clustering algorithms. (3) Cloud model-based C-means clustering methods address the limitations of fuzzy and grey clustering by integrating fuzziness and randomness but suffer from unstable membership degrees and poor result reproducibility due to the randomness of cloud droplet generation; additionally, they lack a set-theoretic framework, making it difficult to handle multi-concept collaborative processing scenarios.

To overcome the aforementioned limitations, this paper proposes a cloud Set-based C-means clustering method and presents its application in the classification of hydrogen energy technology competitive intelligence. The characteristics and innovations of the proposed method are as follows: (1) Strict standard normal cloud sets are used to represent uncertainty, enabling clear and objective expression forms for upper and lower membership degrees. (2) A strict standard normal cloud set and its backward generator are proposed, providing a standard representation method for cloud sets and an effective approach to generate cloud sets from data, laying a solid foundation for C-means clustering. (3) The method and steps of cloud set-based C-means clustering are established, including parameter iterative update formulas and membership degree update methods. (4) The application of this method in the classification of hydrogen energy technology competitive intelligence verifies its effectiveness. The core advantage of the strict standard normal cloud set lies in its membership degree, which possesses both strict orderliness and equidistance—improving the standardization and practicality of the set. The membership degree has clear hierarchical levels, and strict orderliness ensures a distinct gradient of membership strength among elements, allowing direct differentiation of the priority of elements' affiliation to the set. Equidistance ensures that differences in membership strength present a fixed quantitative standard, enhancing the reliability of uncertainty modeling in C-means clustering. (1) Strict ordering Conventional. Classic cloud models suffer from the flaw of core membership deviation, where the core membership value may fall outside the interval bounded by upper and lower

memberships. Strict ordering guarantees that the core membership of strict standard normal cloud sets always resides within the upper–lower membership interval. (2) Area equidistance. Area equidistance constrains the core membership of strict standard normal cloud sets to the central zone of the upper–lower membership interval, and the core membership function curve equally partitions the area of fuzzy regions contained within this interval. Accordingly, the proposed model effectively eliminates the out-of-bounds deviation of core membership and maintains a symmetric, balanced membership distribution structure. It yields more stable and objective quantification results for hybrid fuzziness-randomness uncertainty, fundamentally eradicating one-sided distortion of cluster boundaries and enhancing the consistency and credibility of clustering partition results.

This paper is structured as follows: Firstly, the current status of relevant research is presented in Section 1. Next, the research fundamentals are outlined in Section 2. Subsequently, the strict standard normal cloud set and its inverse generator method are proposed in Section 3. C-means text clustering based on the strict standard normal cloud set is investigated in Section 4. Section 5 presents the application and comparative analysis of this method in the classification of hydrogen energy technology competitive intelligence. Finally, the conclusions are summarized in Section 6.

2. Preliminaries

2.1. Normal Cloud Set

A cloud set is a type-2 uncertain set developed on the basis of the cloud model. A normal cloud set refers to a cloud set whose membership degrees satisfy a normal distribution, which includes the following four parameters: expectation Ex , entropy En , and hyper-entropy He . Among them, expectation Ex is the center of the cloud set where the element has the maximum membership degree. Entropy En represents the degree of uncertainty of the cloud set. Hyper-entropy He , as the entropy of entropy, represents the uncertainty of entropy.

2.2. Membership Degree of the Normal Cloud Set

The membership degree of the normal cloud set is shown in formula (1) [40]:

$$f(x)=\begin{cases} f(x)^U = e^{-\frac{(x-Ex)^2}{2(En+\beta He)^2}} \\ f(x)^M = e^{-\frac{(x-Ex)^2}{2En^2}} \\ f(x)^L = e^{-\frac{(x-Ex)^2}{2(En-\beta He)^2}} \end{cases} \tag{1}$$

where $f(x)^U$ denotes the upper membership degree, $f(x)^L$ denotes the lower membership degree, $f(x)^M$ denotes the middle membership degree, and β is the random coefficient of the cloud set. When $En-\beta He<0$, a truncation of lower membership degree $f(x)^L$ is adopted. Depending on the specific application context, it may be decided whether to retain this formulation or adopt alternative lower membership functions, provided that the effective representation of uncertainty is preserved.

When using the interval of upper and lower membership degrees to express the membership degree of the cloud set, the expression is shown in formula (2) [36-37]:

$$f(x)=[e^{-\frac{(x-Ex)^2}{2(En-\beta He)^2}}, e^{-\frac{(x-Ex)^2}{2(En+\beta He)^2}}] \tag{2}$$

3. Standard Normal Cloud Set and Its Inverse Generator

3.1. Strict Standard Normal Cloud Set

Definition 1: Strict Orderliness of the Membership Degree. If the membership degree of a normal cloud set satisfies the relationship shown in formula (3) below:

$$f(x)^U > f(x)^M > f(x)^L \tag{3}$$

then it is referred to as the strict orderliness of the membership degree.

Definition 2: Equidistance of the Membership Degree. If the membership degree of a normal cloud set satisfies the relationship shown in formula (4) below:

$$f(x)^U - f(x)^M = f(x)^M - f(x)^L \tag{4}$$

then it is said to satisfy the equidistance of the membership degree.

Definition 3: Strict Standard Normal Cloud Set: If a normal cloud set satisfies the following two conditions: strict orderliness of the membership degree and equidistance of the membership degree, it is referred to as a strict standard normal cloud set.

3.2. Backward Generation Method of the Strict Standard Normal Cloud Set

Definition 4: Inverse Generator of the Strict Standard Normal Cloud Set: Given data, it refers to the process of deriving Ex, En, and He and β of the strict standard normal cloud set.

3.2.1. Derivation of *Ex, En* and *He*

Given the data $x=\{x_1,x_2,...,x_n\}$, based on the backward cloud method, the digital characteristics Ex, En and He of the cloud model for x can be obtained using the following formulas (5-7) referring to the relevant literature [41,42]:

$$Ex = 1/n\sum_{i=1}^{n} x_i \tag{5}$$

$$En = \sqrt{\frac{\pi}{2}}\frac{1}{n}\sum_{i=1}^{n}|x_i - Ex| \tag{6}$$

$$\begin{cases} He &= \sqrt{\overline{DY} - \frac{1}{2}\sqrt{\max\left(4\overline{DY}^2 - 2S_{DY}^2\right)}} \\ \overline{DY} &= \frac{1}{n}\sum_{i=1}^{n} DY_i \\ S_{DY}^2 &= \frac{1}{n}\sum_{i=1}^{n}(DY_i - \overline{DY})^2 \\ DY_i &= \frac{1}{m-1}\sum_{j=1}^{m}(x_j^{(i)} - \bar{x}^{(i)})^2 \end{cases} \tag{7}$$

where the mean value is adopt after 1000 random sampling when calculating He. When imaginary numbers occur, that sampling iteration is excluded from the mean calculation.

3.2.2. Derivation of β

(1) β that Satisfies Strict Orderliness

According to the strict orderliness of the membership degree in Definition 1, the following formula (8) holds:

$$e^{-\frac{(x-Ex)^2}{2(En+\beta He)}} > e^{-\frac{(x-Ex)^2}{2(En)}} > e^{-\frac{(x-Ex)^2}{2(En-\beta He)}} \tag{8}$$

Since the exponential function is a strictly monotonic function, the following formulas (9-10) hold:

$$-\frac{(x-Ex)^2}{2(En+\beta He)^2} > -\frac{(x-Ex)^2}{2En^2} > -\frac{(x-Ex)^2}{2(En-\beta He)^2} \tag{9}$$

$$\frac{(x-Ex)^2}{2(En-\beta He)^2} > \frac{(x-Ex)^2}{2En^2} > \frac{(x-Ex)^2}{2(En+\beta He)^2} \tag{10}$$

From the left part of above formula (10), the formula (11-13) can be obtained step by step:

$$\frac{(x-Ex)^2}{2(En-\beta He)^2} > \frac{(x-Ex)^2}{2En^2} \tag{11}$$

$$(En - \beta He)^2 < En^2 \quad (12)$$

$$-2\beta EnHe + (\beta He)^2 < 0 \quad (13)$$

Similarly, from the right part of the above formula (10), formulas (14) can be obtained step by step:

$$En^2 < (En + \beta He)^2 \quad (14)$$

This formula (14) holds identically. Therefore, by combining formulas (13) and (14), we obtain formula (15):

$$0 < \beta < \frac{2En}{He} \quad (15)$$

(2) β that Satisfies the Equidistance of the Membership Degree

According to the equidistance of the membership degree in Definition 2, we have the following formulas (16-17):

$$2f(x)^M = f(x)^L + f(x)^U \quad (16)$$

$$2\, e^{-\frac{(x-Ex)^2}{2En^2}} = e^{-\frac{(x-Ex)^2}{2(En-\beta He)^2}} + e^{-\frac{(x-Ex)^2}{2(En+\beta He)^2}} \quad (17)$$

Since x is a data set, the above formula (17) cannot be directly solved. Replace it with the overall equivalent formula (18) referred to as the difference value, with the following mathematical expression:

$$\int_{-\infty}^{+\infty}(f(x)^U - f(x)^M)^2\, dx = \int_{-\infty}^{+\infty}(f(x)^M - f(x)^L)^2\, dx \quad (18)$$

To solve the above formula (18), numerical integration is used. Combined with formula (15) the final solution of the formula is obtained as shown in formula (19) below:

$$\beta = 1.21\frac{En}{He} \quad (19)$$

where the value 1.21 is the rounded approximation of 1.208198 to two decimal places. A specific truncation is used to represent the model when $En - \beta He < 0$ for $f(x)^L$. When β=1.21$\frac{En}{He}$ negative random entropy is permissible under such circumstances, as the absolute value is used in the actual computation. Here, $E_n - \beta H_e = -0.21E_n$, and $(E_n - \beta H_e)^2 = (-0.21E_n)^2$ is equivalent to truncating the membership function.

4. C-Means Text Clustering Based on Strict Standard Normal Cloud Set

Given n texts with m features and clustering them into h categories, the steps of the entire text clustering based on the strictly standardized normal cloud set are as follows:

Step 1: Construction of feature vectors for Chinese Texts

Word segmentation and cleaning: Segment all texts (e.g., using Jieba for Chinese texts), remove stop words, and punctuation marks, while retaining effective feature words.

Feature extraction: Convert texts into feature vectors using TF-IDF (term frequency–inverse document frequency). Herein, TF refers to Term Frequency, which equals the number of occurrences of a word divided by the total number of words in the document. IDF denotes Inverse Document Frequency, calculated as log(total number of documents in the corpus / (number of documents containing the word + 1)). TF-IDF is the product of TF and IDF (TF-IDF = TF * IDF). The feature vector of the i-th text is denoted as x_i={x_{i1},x_{i2},...,x_{im}}, where x_{ik} represents the TF-IDF value of the k-th feature, thereby forming a feature matrix.

Step 2: Selection of cluster centers

The number of distinct values divided by the number of clouds should be no less than 6, serving as a reference for determining the number of clouds. It is generally not recommended to generate only one cloud, as it may fail to provide sufficient initial candidate clustering centers. Once the number of clouds is determined, the distinct values are evenly partitioned. When the quantity is insufficient, only those partitions containing high-frequency repeated values are allowed to have fewer than 6 distinct values. This method is sensitive to initial clouds, therefore, multiple random initialization runs followed by optimal result selection can be adopted. The mean of each partition is then computed as the peak value point, i.e., the expected value (Ex) of the cloud model. When the number of distinct values is not evenly divisible by the number of clouds, unequal partitioning with a size difference of 1 is permitted. Then, formulas (5-7) are employed to generate the corresponding cloud set parameters. Following this approach: When there is only one peak point: one asymmetric cloud is generated. The interval from 0 to the peak point forms the left ascending cloud, and the interval from the peak point to 1 forms the right descending cloud. When there are multiple peak points: for the leftmost peak point, the left ascending cloud spans from 0 to this peak point, and the right descending cloud spans from this peak point to the midpoint between the first and second peak points; analogously, for the rightmost peak point, the right descending cloud spans from this peak point to 1. The parameters $Ex_{ij,k}$,$En_{ij,k}$,$He_{ij,k}$ and $\beta_{ij,k}$ of the cloud set — corresponding to the k-th peak region, the j-th feature, and the i-th clustering object—are calculated using formulas (5-7) and (19). The cloud sets of each feature may be asymmetric or half-cloud, where each asymmetric cloud consists of a left half-cloud and a right half-cloud. Its parameters are recorded as $(E_{x_{ij,k}}, E^l_{n_{ij,k}}, H^l_{e_{ij,k}}, E^r_{n_{ij,k}}, H^r_{e_{ij,k}}, \beta^l_{ij,k}, \beta^r_{ij,k})$.

Candidate cluster centers are formed by the Cartesian product of the cloud sets corresponding to all features. First, compute the lower membership degree of each feature value for each text and candidate center. For the n existing clustered text objects, if all feature dimensions satisfy the following matching condition, the text is considered to match the candidate center: the corresponding degree exceeds the threshold ε defined in Formula (20). If a text matches multiple centers, its final affiliation is determined by "the maximum sum of lower membership degrees across multiple features".

$$exp(-(x_{ij} - Ex_{ij,k})^2/(2 * (En_{ij,k} - \beta_{ij,k}He_{ij,k})^2)) >= \varepsilon \quad (20)$$

where i=1, 2, ..., n , j=1, 2, ..., m.

Record whether a text matches a candidate cluster center by formula (21) as follow: If formula (20) is satisfied and the sum of lower membership degrees is maximum, then $n_{x_{i,k}}$ is equal to 1; therwise, $n_{x_{i,k}}$ is equal to 0.

$$n_{x_{i,k}} = \begin{cases} 1, & if\ \epsilon^{-\frac{-(x_{ij}-Ex_{ij,k})^2}{2*(En_{ij,k}-\beta_{ij,k}He_{ij,k})^2}} >= \varepsilon \quad and\ argmax \sum_{j=1}^{m} e^{-\frac{-(x_{ij}-Ex_{ij,k})^2}{2*(En_{ij,k}-\beta_{ij,k}He_{ij,k})^2}}\ for\ all\ i,j \\ 0, & if\quad e^{-\frac{-(x_{ij}-Ex_{ij,k})^2}{2*(En_{ij,k}-\beta_{ij,k}He_{ij,k})^2}} < \varepsilon\ for\ all\ i,j \end{cases} \quad (21)$$

Iterate over all clustering objects and calculate the number $N_{mat,k}$ of matching samples for each candidate cluster center by formula (22):

$$N_{mat,k} = \sum_{i=1}^{n} n_{x_{i,k}} \quad (22)$$

Candidate cluster centers with the number of matching samples $N_{mat,k}$ exceeding the predefined threshold are

chosen as initial cluster centers.

Step 3: Initialize the cloud membership degree matrix

For the *k*-th feature of the *j*-th cluster in clustering, the parameters Ex_{jk}, En_{jk}, He_{jk}, and β_{jk} of the initial inverse standard normal cloud set are calculated by formulas (5-7) and (19). The initial upper membership degree $f(x_{ik})^U{}_{i\in j}$ and initial lower membership degree $f(x_{ik})^L{}_{i\in j}$ for sample *i* belonging to the *j*-th cluster(denoted as $i\in j$) in terms of the *k*-th feature are obtained through calculation using formula (1). Then, the initial comprehensive membership degree matrix $u_{n,m}$ is derived as follow in formula (23):

$$u_{n,m}=\begin{bmatrix} [f(x_1)^L{}_{1\in1}, f(x_1)^U{}_{1\in j}] & [f(x_1)^L{}_{1\in2}, f(x_1)^U{}_{1\in2}] & \cdots & [f(x_1)^L{}_{1\in h}, f(x_1)^U{}_{1\in h}] \\ [f(x_2)^L{}_{2\in1}, f(x_2)^U{}_{2\in j}] & [f(x_2)^L{}_{2\in2}, f(x_2)^U{}_{2\in2}] & \cdots & [f(x_2)^L{}_{2\in h}, f(x_2)^U{}_{2\in h}]. \\ \vdots & \vdots & \cdots & \vdots \\ [f(x_n)^L{}_{n\in1}, f(x_n)^U{}_{n\in1}] & [f(x_n)^L{}_{n\in2}, f(x_n)^U{}_{n\in2}] & \cdots & [f(x_n)^L{}_{n\in h}, f(x_n)^U{}_{n\in h}] \end{bmatrix} \quad (23)$$

Normalize the initial comprehensive membership degree matrix by formulas (24-25) below.

$$f(x_{ik})^L_{nor}=f(x_{ik})^L/\sum_{\substack{j=1\\ j\neq i}}^{m} f(x_{jk})^L \quad (24)$$

$$f(x_{ik})^U_{nor}=f(x_{ik})^U/\sum_{\substack{j=1\\ j\neq i}}^{m} f(x_{jk})^U \quad (25)$$

The final comprehensive upper and lower membership degrees for *i*-th sample belonging to *j*-th class are equal to the product of the membership degrees of each feature, as shown in formulas (26) and (27) below:

$$f(x_i)^U{}_{total}=\prod_{k=1}^{m} f(x_{ik})^L_{nor} \quad (26)$$

$$f(x_i)^L{}_{total}=\prod_{k=1}^{m} f(x_{ik})^U_{nor} \quad (27)$$

Unlike the initially randomly assigned membership degrees in fuzzy C-means clustering, combining them with the membership degrees of the strict standard normal cloud set derived from data yields the following benefits: enhancing the quality of clustering initialization and the stability of results, while reducing iterative costs and improving the interpretability of outcomes.

Step 4: Iterative update

Define the objective function as "the sum of squared distances of the parameters at the *t*-th iteration, weighted by the *m*-th power of the membership degrees at the (*t*-1)-th iteration" as shown in following formula (28):

$$J^{(t)}=\sum_{i=1}^{n} {u_i}^{(t-1)^m}(x_i-Ex^{(t)})^2 \quad (28)$$

According to formula (1), the membership degrees at the *t*-th iteration are calculated by the following formula (29):

$${u_i}^{(t-1)}=[e^{-\frac{(x_i-Ex^{(t-1)})^2}{2(En^{(t-1)}-\beta^{(t-1)}He^{(t-1)})^2}}, e^{-\frac{(x_i-Ex^{(t-1)})^2}{2(En^{(t-1)}+\beta^{(t-1)}He^{(t-1)})^2}}] \quad (29)$$

(1)Iterative formula for Ex

Take the partial derivative of $J^{(t)}$ with respect to Ex; since ${u_i}^{(t-1)}$ is independent of $Ex^{(t)}$, we obtain formula (30):

$$\frac{\partial J^{(t)}}{\partial Ex^{(t)}}=\sum_{i=1}^{n} {u_i}^{(t-1)^m} 2(x_i-Ex^{(t)})(-1)=0 \quad (30)$$

where *m* is the fuzziness factor, which is generally set to *m* = 2.

Solve formula (41), and we yield the iterative update formula for $Ex^{(t)}$, i.e., formula (31):

$$Ex^{(t)}=\frac{\sum_{i=1}^{n} {u_i}^{(t-1)^m} x_i}{\sum_{i=1}^{n} {u_i}^{(t-1)^m}} \quad (31)$$

(2)Iterative formula for En

Take the partial derivative of $J^{(t)}$ with respect to En, set it to zero, we obtain formula (32):

$$\frac{\partial J^{(t)}}{\partial En^{(t-1)}}=\sum_{i=1}^{n} m{u_i}^{(t-1)^m}\frac{(x_i-Ex^{(t-1)})^2}{(\sqrt{\sum_{i=1}^{n} {u_i}^{(t-1)^m}(x_i-Ex^{(t-1)})^2})^3}(x_i-Ex^{(t)})^2=0 \quad (32)$$

This partial derivative formula in formula (30) has no solution. An alternative method is used to calculate the iterative formula for En. Based on the requirement that the width term at the *t*-th iteration should reflect the degree of dispersion between the new center $Ex^{(t)}$ and the data set, together with the statistical significance of minimizing the objective function, we derive the following formula (33):

$$(En^{(t)}\pm\beta^{(t)}He^{(t)})^2=\frac{\sum_{i=1}^{n} {u_i}^{(t-1)^m}(x_i-Ex^{(t)})^2}{\sum_{i=1}^{n} {u_i}^{(t-1)^m}} \quad (33)$$

From formula (33), different En values for the upper and lower membership degree scenarios can be derived, as presented in formula (34):

$$En^{(t)}=\begin{cases}\sqrt{\frac{\sum_{i=1}^{n} {u_i}^{(t-1)^m}(x_i-Ex^{(t)})^2}{\sum_{i=1}^{n} {u_i}^{(t-1)^m}}}/2.21 \ (upper)\\ \sqrt{\frac{\sum_{i=1}^{n} {u_i}^{(t-1)^m}(x_i-Ex^{(t)})^2}{\sum_{i=1}^{n} {u_i}^{(t-1)^m}}}/0.21 \ (lower)\end{cases} \quad (34)$$

(3)Iterative formula for He

Similarly, take the partial derivative of $J^{(t)}$ with respect to He, and set it to 0. It again has no solution. Based on the same requirement that the width at the *t*-th iteration should reflect the degree of dispersion between the new center $En^{(t)}$ and the data, we obtain formula (35):

$$(He^{(t)})^2=\frac{\sum_{i=1}^{n} {u_i}^{(t-1)^m}(|x_i-Ex^{(t)}|-En^{(t)})^2}{\sum_{i=1}^{n} {u_i}^{(t-1)^m}} \quad (35)$$

(4)Iterative formula for β

The iterative formula for β is shown in formula (36) below:

$$\beta^{(t)}=\frac{1.21En^{(t)}}{He^{(t)}} \quad (36)$$

The new upper and lower membership degrees for the single feature are calculated by formulas (37) below:

$$\begin{cases} f(x_i)^U{}_{i\in j}{}^{(t)}=e^{-\frac{(x_{ik}-Ex_{jk}{}^{(t)})^2}{2(En_{jk}{}^{(t)}+\beta_{jk}{}^{(t)}He_{jk}{}^{(t)})^2}} \\ f(x_i)^M{}_{i\in j}{}^{(t)}=e^{-\frac{(x_{ik}-Ex_{jk}{}^{(t)})^2}{2(En_{jk}{}^{(t)})^2}} \\ f(x_i)^L{}_{i\in j}{}^{(t)}=e^{-\frac{(x_{ik}-Ex_{jk}{}^{(t)})^2}{2(En_{jk}{}^{(t)}-\beta_{jk}{}^{(t)}He_{jk})^2}} \end{cases} \quad (37)$$

The overall upper and lower membership degrees of sample *i* belonging to class *j* are calculated as the product of the membership degrees of each feature, as shown in formula (38) below:

$$\begin{cases} f(x_i)^U{}_{i\in j}{}^{(t)}=\prod_{k=1}^{m} f(x_{ik})^U{}_{i\in j}{}^{(t)}=\prod_{k=1}^{m} e^{-\frac{(x_{ik}-Ex_{jk}{}^{(t)})^2}{2(En_{jk}{}^{(t)}+\beta_{jk}{}^{(t)}He_{jk}{}^{(t)})^2}} \\ f(x_i)^M{}_{i\in j}{}^{(t)}=\prod_{k=1}^{m} f(x_{ik})^M{}_{i\in j}{}^{(t)}=\prod_{k=1}^{m} e^{-\frac{(x_{ik}-Ex_{jk}{}^{(t)})^2}{2(En_{jk}{}^{(t)})^2}} \\ f(x_i)^L{}_{i\in j}{}^{(t)}=\prod_{k=1}^{m} f(x_{ik})^L{}_{i\in j}{}^{(t)}=\prod_{k=1}^{m} e^{-\frac{(x_{ik}-Ex_{jk}{}^{(t)})^2}{2(En_{jk}{}^{(t)}-\beta_{jk}{}^{(t)}He_{jk})^2}} \end{cases} \quad (38)$$

Perform the membership degree normalization using formula (39) below:

$$\begin{cases} f(x_i)^U{}_{i\in j}{}^{(t)} = \dfrac{f(x_i)^U{}_{i\in j}{}^{(t)}}{\sum_{k=1}^{h} f(x_i)^U{}_{i\in k}{}^{(t)}} \ (j = 1,2,\ldots,h) \\ f(x_i)^M{}_{i\in j}{}^{(t)} = \dfrac{f(x_i)^M{}_{i\in j}{}^{(t)}}{\sum_{k=1}^{h} f(x_i)^M{}_{i\in k}{}^{(t)}} \ (j = 1,2,\ldots,h) \\ f(x_i)^L{}_{i\in j}{}^{(t)} = \dfrac{f(x_i)^L{}_{i\in j}{}^{(t)}}{\sum_{k=1}^{h} f(x_i)^L{}_{i\in k}{}^{(t)}} \ (j = 1,2,\ldots,h) \end{cases} \quad (39)$$

Update the comprehensive membership degree matrix, as shown in formula (40) below:

$$u_{nm}{}^{(t)} = \begin{bmatrix} [f(x_1)^L{}_{1\in 1}{}^{(t)}, f(x_1)^M{}_{1\in 1}{}^{(t)}, f(x_1)^U{}_{1\in 1}{}^{(t)}] & [f(x_1)^L{}_{1\in 2}{}^{(t)}, f(x_1)^M{}_{1\in 2}{}^{(t)}, f(x_1)^U{}_{1\in 2}{}^{(t)}] & \cdots & [f(x_1)^L{}_{1\in h}{}^{(t)}, f(x_1)^M{}_{1\in h}{}^{(t)}, f(x_1)^U{}_{1\in h}{}^{(t)}] \\ [f(x_2)^L{}_{2\in 1}{}^{(t)}, f(x_2)^M{}_{2\in 1}{}^{(t)}, f(x_2)^U{}_{2\in 1}{}^{(t)}] & [f(x_2)^L{}_{2\in 2}{}^{(t)}, f(x_2)^M{}_{2\in 2}{}^{(t)}, f(x_2)^U{}_{2\in 2}{}^{(t)}] & \cdots & [f(x_2)^L{}_{2\in h}{}^{(t)}, f(x_2)^M{}_{2\in h}{}^{(t)}, f(x_2)^U{}_{2\in h}{}^{(t)}] \\ \vdots & \vdots & \cdots & \vdots \\ [f(x_n)^L{}_{n\in 1}{}^{(t)}, f(x_n)^M{}_{n\in 1}{}^{(t)}, f(x_n)^U{}_{n\in 1}{}^{(t)}] & [f(x_n)^L{}_{n\in 2}{}^{(t)}, f(x_n)^M{}_{n\in 2}{}^{(t)}, f(x_n)^U{}_{n\in 2}{}^{(t)}] & \cdots & [f(x_n)^L{}_{n\in h}{}^{(t)}, f(x_n)^M{}_{n\in h}{}^{(t)}, f(x_n)^U{}_{n\in h}{}^{(t)}] \end{bmatrix} \quad (40)$$

Calculate the objective function $K^{(t)}$ in formula (41), which is the sum of the absolute values of the differences between the upper and lower membership degrees at all corresponding positions at the t-th and (t-1)-th iterations:

$$\begin{cases} \Delta_U(t) = \left\| f(x_i)^U{}_{i\in j}{}^{(t)} - f(x_i)^U{}_{i\in j}{}^{(t-1)} \right\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{h}(f(x_i)^U{}_{i\in j}{}^{(t)} - f(x_i)^U{}_{i\in j}{}^{(t-1)})^2} \\ \Delta_M(t) = \left\| f(x_i)^M{}_{i\in j}{}^{(t)} - f(x_i)^M{}_{i\in j}{}^{(t-1)} \right\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{h}(f(x_i)^M{}_{i\in j}{}^{(t)} - f(x_i)^M{}_{i\in j}{}^{(t-1)})^2} \\ \Delta_L(t) = \left\| f(x_i)^L{}_{i\in j}{}^{(t)} - f(x_i)^L{}_{i\in j}{}^{(t-1)} \right\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{h}(f(x_i)^L{}_{i\in j}{}^{(t)} - f(x_i)^L{}_{i\in j}{}^{(t-1)})^2} \end{cases} \quad (41)$$

Let ε be the threshold, if formula (42) below is satisfied:

$$\begin{cases} \Delta_U(t) \le \varepsilon \\ \Delta_M(t) \le \varepsilon \\ \Delta_L(t) \le \varepsilon \end{cases} \quad (42)$$

where the preset convergence threshold ε=0.0001. If the iteration differences reach the threshold, then it is determined to have converged. Terminate the iteration, and proceed to the next step; otherwise, return to Step 3.

STEP 5: Output the clustering results.

Each text i is assigned to the class with the maximum average membership degree as shown in formula (43) below:

$$argmax_{j=1,2,\ldots,h} \frac{(f(x_i)^L{}_{i\in j}{}^{(t)} + (f(x_i)^M{}_{i\in j}{}^{(t)} + f(x_i)^U{}_{i\in j}{}^{(t)})}{3} \quad (43)$$

Alternatively, one or more of the following alternative metrics in formulas (44--47) could be adopted to assign the samples.

$$f(x_i)^U{}_{i\in j}{}^{(t)} > \alpha_1 \quad (44)$$

$$f(x_i)^M{}_{i\in j}{}^{(t)} > \alpha_2 \quad (45)$$

$$f(x_i)^L{}_{i\in j}{}^{(t)} > \alpha_3 \quad (46)$$

$$\frac{(f(x_i)^L{}_{i\in j}{}^{(t)} + (f(x_i)^M{}_{i\in j}{}^{(t)} + f(x_i)^U{}_{i\in j}{}^{(t)})}{3} > \alpha_4 \quad (47)$$

where $\alpha_i (i = 1,2,3,4)$ is the threshold value.

5. Application in the Competitive Analysis of International Scientific and Technological Intelligence

5.1. Application of Standard Normal Cloud Set C-means Clustering

Table 1. Feature vectors of texts

ID	Feature vector	ID	Feature vector	ID	Feature vector
1	[0.00, 0.00, 0.00, 0.20, 0.39]	2	[0.69, 0.00, 0.00, 0.00, 0.19]	3	[0.00, 0.66, 0.00, 0.00, 0.00]
4	[0.44, 0.00, 1.00, 0.00, 0.00]	5	[0.00, 0.00, 0.00, 0.23, 0.00]	6	[0.00, 0.00, 0.00, 0.00, 0.79]
7	[1.00, 0.00, 0.00, 0.00, 0.00]	8	[0.00, 0.13, 0.00, 0.00, 0.00]	9	[0.00, 0.00, 0.58, 0.00, 0.00]
10	[0.00, 0.00, 0.00, 0.87, 0.00]	11	[0.00, 0.00, 0.00, 0.00, 0.64]	12	[0.24, 0.00, 0.00, 0.00, 0.00]
13	[0.00, 0.48, 0.00, 0.00, 0.00]	14	[0.00, 0.00, 0.21, 0.00, 0.00]	15	[0.00, 0.00, 0.00, 0.66, 0.00]
16	[0.00, 0.00, 0.00, 0.20, 0.79]	17	[0.50, 0.00, 0.00, 0.00, 0.00]	18	[0.00, 0.50, 0.00, 0.00, 0.00]
19	[0.00, 0.00, 0.60, 0.21, 0.00]	20	[0.00, 0.00, 0.00, 0.25, 0.00]	21	[0.00, 0.00, 0.00, 0.00, 0.49]
22	[0.18, 0.00, 0.22, 0.15, 0.00]	23	[0.00, 1.00, 0.00, 0.00, 0.00]	24	[0.00, 0.00, 0.33, 0.19, 0.18]
25	[0.00, 0.00, 0.00, 0.25, 0.00]	26	[0.00, 0.00, 0.00, 0.23, 1.00]	27	[0.92, 0.00, 0.00, 0.00, 0.00]
28	[0.00, 0.44, 0.00, 0.00, 0.00]	29	[0.00, 0.00, 0.21, 0.29, 0.00]	30	[0.00, 0.00, 0.00, 1.00, 0.00]
31	[0.00, 0.00, 0.00, 0.19, 0.19]	32	[0.89, 0.00, 0.00, 0.00, 0.00]	33	[0.00, 0.00, 0.62, 0.00, 0.00]
34	[0.00, 0.00, 0.00, 0.21, 0.00]	35	[0.00, 0.00, 0.00, 0.19, 0.47]	36	[0.54, 0.00, 0.00, 0.00, 0.00]
37	[0.00, 0.72, 0.00, 0.00, 0.00]	38	[0.00, 0.00, 0.11, 0.00, 0.00]	39	[0.00, 0.00, 0.00, 0.48, 0.00]
40	[0.00, 0.00, 0.00, 0.18, 0.35]	41	[0.16, 0.00, 0.19, 0.13, 0.26]	42	[0.44, 0.00, 0.00, 0.00, 0.00]
43	[0.00, 0.16, 0.00, 0.22, 0.00]	44	[0.00, 0.00, 0.71, 0.00, 0.00]	45	[0.00, 0.00, 0.00, 0.91, 0.00]
46	[0.00, 0.00, 0.00, 0.20, 0.20]	47	[0.78, 0.00, 0.00, 0.00, 0.00]	48	[0.00, 0.24, 0.00, 0.00, 0.00]
49	[0.00, 0.00, 0.31, 0.00, 0.00]	50	[0.00, 0.00, 0.00, 0.45, 0.18]	51	[0.00, 0.00, 0.00, 0.00, 0.59]
52	[0.23, 0.00, 0.00, 0.00, 0.00]	53	[0.00, 0.16, 0.00, 0.00, 0.00]	54	[0.00, 0.00, 0.13, 0.36, 0.09]
55	[0.00, 0.00, 0.00, 0.20, 0.39]	56	[0.00, 0.00, 0.00, 0.00, 0.47]	57	[0.28, 0.00, 0.00, 0.00, 0.00]
58	[0.00, 0.20, 0.00, 0.28, 0.00]	59	[0.00, 0.00, 0.39, 0.00, 0.00]	60	[0.00, 0.00, 0.00, 0.40, 0.00]

(1)Feature vector extraction of texts

Sixty Chinese texts related to the competitive analysis of international scientific and technological intelligence on hydrogen energy were selected and collected via the Internet.

All collected textual data are classified into six predefined thematic clusters, with 11 for hydrogen production, 11 for hydrogen technology, 10 for hydrogen cooperation, 12 for hydrogen infrastructure, 12 for hydrogen policy and 4 for mixed themes that do not belong to any of the above categories. These texts were processed using Jieba, a Python-based Chinese word segmentation library, followed by stop-word removal to obtain feature terms. Through term frequency filtering and TF-IDF (Term Frequency-Inverse Document Frequency) transformation, the feature vectors shown in Table 1 were generated, which include five key features: green hydrogen production capacity (F1), fuel cell technology (F2), international cooperation (F3), hydrogen energy infrastructure (F4), and subside policies (F5). Each feature contains 30 feature word groups. Consequently, in this case, each text class is mainly associated with a single feature word group.

Table 2. Peak Value and corresponding cloud sets

Feature	Peak Positions	Number of cloud	Cloud Sets (Ex, En_L, En_R, He_L, He_R, β_L, β_R)
F1	0.22, 0.76	2	C1-1(0.22,0.27,0.13,0.04,0.02,8.17,7.86); C1-2(0.76,0.23,0.17,0.03,0.03,9.28,6.86)
F2	0.15, 0.63	2	C2-1(0.15,0.18,0.06,0.00,0.01,0.00,7.26); C2-2(0.63,0.20,0.20,0.01,0.05,24.20,4.84)
F3	0.17, 0.60	2	C3-1(0.17,0.20,0.12,0.03,0.02,8.07,7.26); C3-2(0.60,0.10,0.22,0.03,0.06,4.03,4.44)
F4	0.18, 0.57	2	C4-1(0.18,0.21,0.06,0.03,0.02,8.47,3.63); C4-2(0.57,0.16,0.36,0.01,0.04,19.36,10.89)
F5	0.18, 0.62	2	C5-1(0.18,0.21,0.13,0.04,0.02,6.35,7.86); C5-2(0.62,0.15,0.23,0.02,0.04,9.07,6.96)

(2)Selection of clustering centers

Using Step 1 and Step 2 in Section 4 “C-Means Text Clustering Based on Strict Standard Normal Cloud Set”, for the 60 values of each feature, the numbers of cloud are selected as 2, 2, 2, and 2 (This method is sensitive to initial clouds, so explorations and comparisons can be conducted to determine the optimal number of cloud.), their corresponding cloud sets were calculated , as shown in Table 2.

Based on the number of dense regions and their corresponding cloud sets for each feature, the cloud sets corresponding to the dense regions of F1-F5 are subjected to "Cartesian Product Combination" to generate all possible candidate clustering centers. A total of 32 candidate clustering centers (2×2×2×2×2) are generated, namely C01 (C1-1, C2-1, C3-1, C4-1, C5-1), C02 (C1-1, C2-1, C3-1, C4-1, C5-2), C03 (C1-1, C2-1, C3-1, C4-2, C5-1), ..., C32 (C1-2, C2-2, C3-2, C4-2, C5-2).

The Text Matching Threshold was set to 3. Candidate clustering centers with a matched text count greater than 3 were selected as formal initial clustering centers, with the resulting 6 formal initial clustering centers shown in Table 3 below:

Table 3. Initial clustering centers

ID	Matched Text Count	Mean of upper membership degree	Matched Text Serial Number
C01	27	0.90	1 5 8 12 14 17 20 22 24 25 29 31 34 38 40 41 42 43 46 48 49 52 53 55 57 58 59
C02	8	0.89	6, 11, 16, 21, 26, 35, 51, 56
C03	8	0.89	10, 15, 30, 39, 45, 50, 54, 60
C05	5	0.88	4, 9, 19, 33, 44
C09	6	0.88	3, 13, 18, 23, 28, 37
C17	6	0.89	2, 7, 27, 32, 36, 47

Among them, the cloud set parameters of the initial clustering centers are shown in Table 4 below:

Table 4. Cloud set parameters of clustering centers

Clustering Center	F1 cloud set(Ex,En,He,β)	F2 cloud set(Ex,En,He,β)	F3 cloud set(Ex,En,He,β)	F4 cloud set(Ex,En,He,β)	F5 cloud set(Ex,En,He,β)
C01	C1-1(0.22,0.20,0.03,8.07)	C2-1(0.15,0.12,0.01,14.52)	C3-1(0.17,0.16,0.03,6.45)	C4-1(0.18,0.14,0.03,5.65)	C5-1(0.18,0.17,0.03,6.86)
C02	C1-1(0.22,0.20,0.03,8.07)	C2-1(0.15,0.12,0.01,14.52)	C3-1(0.17,0.16,0.03,6.45)	C4-1(0.18,0.14,0.03,5.65)	C5-2(0.62,0.19,0.03,7.66)
C03	C1-1(0.22,0.20,0.03,8.07)	C2-1(0.15,0.12,0.01,14.52)	C3-1(0.17,0.16,0.03,6.45)	C4-2(0.57,0.26,0.03,10.49)	C5-1(0.18,0.17,0.03,6.86)
C05	C1-1(0.22,0.20,0.03,8.07)	C2-1(0.15,0.12,0.01,14.52)	C3-2(0.60,0.16,0.05,3.87)	C4-1(0.18,0.14,0.03,5.65)	C5-1(0.18,0.17,0.03,6.86)
C09	C1-1(0.22,0.20,0.03,8.07)	C2-2(0.63,0.20,0.03,8.07)	C3-1(0.17,0.16,0.03,6.45)	C4-1(0.18,0.14,0.03,5.65)	C5-1(0.18,0.17,0.03,6.86)
C17	C1-2(0.76,0.20,0.03,8.07)	C2-1(0.15,0.12,0.01,14.52)	C3-1(0.17,0.16,0.03,6.45)	C4-1(0.18,0.14,0.03,5.65)	C5-1(0.18,0.17,0.03,6.86)

(3)Initialization of membership degree matrix

Sixty values of the "green hydrogen production capacity" feature dimension were extracted from the previous feature vector (corresponding to Text Serial Numbers 1-60):[0.00, 0.69, 0.00, 0.44, 0.00, 0.00, 1.00, 0.00, 0.00, 0.00, 0.00, 0.24, 0.00, 0.00, 0.00, 0.00, 0.50, 0.00, 0.00, 0.00, 0.00, 0.18, 0.00, 0.00, 0.00, 0.00, 0.92, 0.00, 0.00, 0.00, 0.00, 0.89, 0.00, 0.00, 0.00, 0.54, 0.00, 0.00, 0.00, 0.00, 0.16, 0.44, 0.00, 0.00, 0.00, 0.00, 0.78, 0.00, 0.00, 0.00, 0.00, 0.23, 0.00, 0.00, 0.00, 0.00, 0.28, 0.00, 0.00, 0.00]. Based on the reverse cloud set modeling method proposed in formulas (5-7) and (19), the following results were obtained:

$$Ex_1 = 0.12$$
$$En_1 = 0.23$$
$$He_1 = 0.12$$
$$\beta = 2.32$$

Using the same method, the cloud set parameters of the other four features were obtained, as shown in Table 5 below:

The data in Table 5 were substituted into formula (1) to yield the upper and lower membership functions of the following five features, as shown in Table 6. The corresponding curves are presented in Figure 1:

The feature vector values of the 60 texts (see Table 1) were substituted into the upper and lower membership functions of the cloud sets of the five features, yielding the initial upper and lower membership degree intervals of the features as shown in Table 7 below.

Table 5. Cloud set parameters of five features

Feature	*Ex*	*En*	*He*	*β*
F1	0.12	0.23	0.12	2.32
F2	0.08	0.16	0.08	2.42
F3	0.09	0.18	0.09	2.42
F4	0.15	0.21	0.11	2.31
F5	0.13	0.23	0.11	2.53

Table 6. Membership functions of the following five features

Feature	Upper membership degree	Medium membership degree	Lower membership degree
F1	exp[- (x - 0.12)² / 2*(0.5084)²]	exp[- (x - 0.12)² / 2*(0.2300)²]	exp[- (x - 0.12)² / 2*(0.0484)²]
F2	exp[- (x - 0.08)² / 2*(0.3536)²]	exp[- (x - 0.08)² / 2*(0.1600)²]	exp[- (x - 0.08)² / 2*(0.0336)²]
F3	exp[- (x - 0.09)² / 2*(0.3978)²]	exp[- (x - 0.09)² / 2*(0.1800)²]	exp[- (x - 0.09)² / 2*(0.0378)²]
F4	exp[- (x - 0.15)² / 2*(0.4641)²]	exp[- (x - 0.15)² / 2*(0.2100)²]	exp[- (x - 0.15)² / 2*(0.0441)²]
F5	exp[- (x - 0.13)² / 2*(0.5083)²]	exp[- (x - 0.13)² / 2*(0.2300)²]	exp[- (x - 0.13)² / 2*(0.0483)²]

The upper and lower membership degrees in Table 7 were normalized respectively, yielding the normalized membership degrees in Table 8 (Rounding errors in the summation of normalized membership degrees can be neglected.). Note that

in Table 8, there is no strict constraint that the normalized upper membership degree is greater than the normalized lower membership degree, so these values no longer represent valid membership degree intervals.

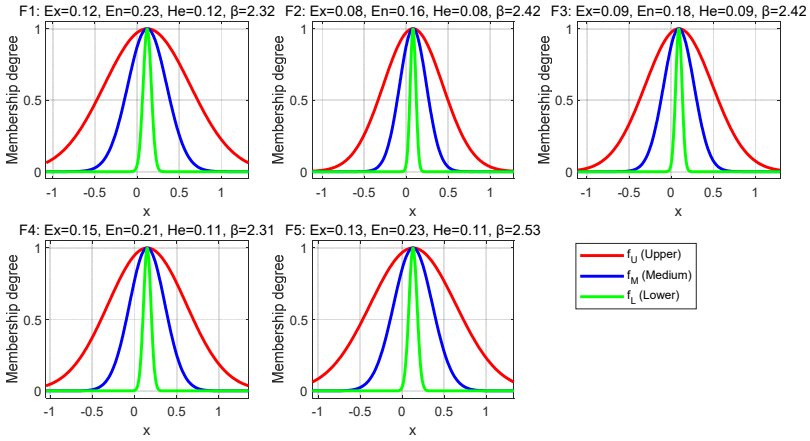


Figure 1. Cloud sets for five features

Table 7. Initial upper, medium and lower membership degree intervals of features

ID	Upper, Medium and Lower Membership Degree(Five features)	ID	Upper, Medium and Lower Membership Degree(Five features)
1	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.52,0.97,0.99],[0.00,0.53,0.88]	31	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.66,0.98,1.00],[0.46,0.97,0.99]
2	[0.00,0.05,0.53],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.46,0.97,0.99]	32	[0.00,0.00,0.32],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]
3	[0.05,0.87,0.97],[0.00,0.00,0.15],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]	33	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.00,0.01,0.41],[0.00,0.77,0.95],[0.03,0.85,0.97]
4	[0.00,0.38,0.82],[0.66,0.75,0.96],[0.00,0.00,0.07],[0.00,0.77,0.95],[0.03,0.85,0.97]	34	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.39,0.96,0.99],[0.03,0.85,0.97]
5	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.18,0.93,0.99],[0.03,0.85,0.97]	35	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.66,0.98,1.00],[0.00,0.34,0.80]
6	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.00,0.02,0.43]	36	[0.00,0.19,0.71],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]
7	[0.00,0.00,0.22],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]	37	[0.05,0.87,0.97],[0.00,0.00,0.10],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]
8	[0.05,0.87,0.97],[0.85,0.89,0.99],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]	38	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.87,0.99,1.00],[0.00,0.77,0.95],[0.03,0.85,0.97]
9	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.00,0.02,0.47],[0.00,0.77,0.95],[0.03,0.85,0.97]	39	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.29,0.78],[0.03,0.85,0.97]
10	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.00,0.30],[0.03,0.85,0.97]	40	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.79,0.99,1.00],[0.00,0.63,0.91]
11	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.00,0.09,0.60]	41	[0.71,0.98,1.00],[0.66,0.75,0.96],[0.03,0.86,0.97],[0.90,1.00,1.00],[0.03,0.85,0.97]
12	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]	42	[0.00,0.38,0.82],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]
13	[0.05,0.87,0.97],[0.00,0.00,0.41],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]	43	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.27,0.95,0.99],[0.03,0.85,0.97]
14	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.01,0.80,0.96],[0.00,0.77,0.95],[0.03,0.85,0.97]	44	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.00,0.00,0.30],[0.00,0.77,0.95],[0.03,0.85,0.97]
15	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.05,0.55],[0.03,0.85,0.97]	45	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.00,0.26],[0.03,0.85,0.97]
16	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.52,0.97,0.99],[0.00,0.02,0.43]	46	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.52,0.97,0.99],[0.35,0.95,0.99]
17	[0.00,0.26,0.76],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]	47	[0.00,0.02,0.43],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]
18	[0.05,0.87,0.97],[0.00,0.00,0.37],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]	48	[0.05,0.87,0.97],[0.19,0.32,0.87],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]
19	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.00,0.02,0.44],[0.39,0.96,0.99],[0.03,0.85,0.97]	49	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.00,0.47,0.86],[0.00,0.77,0.95],[0.03,0.85,0.97]
20	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.07,0.89,0.98],[0.03,0.85,0.97]	50	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.36,0.81],[0.58,0.98,1.00]
21	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.00,0.29,0.78]	51	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.00,0.14,0.66]
22	[0.46,0.97,0.99],[0.66,0.75,0.96],[0.00,0.77,0.95],[1.00,1.00,1.00],[0.03,0.85,0.97]	52	[0.08,0.89,0.98],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]
23	[0.05,0.87,0.97],[0.00,0.00,0.01],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]	53	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]
24	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.00,0.41,0.83],[0.66,0.98,1.00],[0.58,0.98,1.00]	54	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.57,0.98,0.99],[0.00,0.61,0.90],[0.71,0.98,1.00]
25	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.07,0.89,0.98],[0.03,0.85,0.97]	55	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.52,0.97,0.99],[0.00,0.53,0.88]
26	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.18,0.93,0.99],[0.00,0.00,0.23]	56	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.00,0.34,0.80]
27	[0.00,0.00,0.29],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]	57	[0.00,0.79,0.95],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]
28	[0.05,0.87,0.97],[0.00,0.00,0.49],[0.06,0.88,0.97],[0.00,0.77,0.95],[0.03,0.85,0.97]	58	[0.05,0.87,0.97],[0.39,0.53,0.92],[0.06,0.88,0.97],[0.01,0.83,0.96],[0.03,0.85,0.97]
29	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.01,0.80,0.96],[0.01,0.80,0.96],[0.03,0.85,0.97]	59	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.00,0.25,0.75],[0.00,0.77,0.95],[0.03,0.85,0.97]
30	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.00,0.19],[0.03,0.85,0.97]	60	[0.05,0.87,0.97],[0.66,0.75,0.96],[0.06,0.88,0.97],[0.00,0.49,0.86],[0.03,0.85,0.97]

Table 8. Initial normalized membership degrees

ID	Normalized(L,M,U)	ID	Normalized(L,M,U)
1	[0.04,0.22,0.20],[0.51,0.19,0.20],[0.05,0.22,0.20],[0.40,0.24,0.21],[0.00,0.13,0.18]	31	[0.03,0.20,0.20],[0.35,0.17,0.20],[0.03,0.20,0.20],[0.35,0.22,0.20],[0.24,0.22,0.20]
2	[0.00,0.01,0.12],[0.56,0.22,0.22],[0.05,0.26,0.22],[0.00,0.23,0.22],[0.39,0.28,0.23]	32	[0.00,0.00,0.08],[0.88,0.23,0.23],[0.08,0.27,0.23],[0.00,0.24,0.23],[0.04,0.26,0.23]
3	[0.36,0.26,0.24],[0.00,0.00,0.04],[0.43,0.26,0.24],[0.00,0.23,0.24],[0.21,0.25,0.24]	33	[0.07,0.27,0.23],[0.89,0.23,0.23],[0.00,0.00,0.10],[0.00,0.24,0.22],[0.04,0.26,0.23]
4	[0.00,0.14,0.22],[0.96,0.27,0.25],[0.00,0.00,0.02],[0.00,0.28,0.25],[0.04,0.31,0.26]	34	[0.04,0.20,0.20],[0.55,0.17,0.20],[0.05,0.20,0.20],[0.33,0.22,0.20],[0.03,0.20,0.20]
5	[0.05,0.20,0.20],[0.67,0.18,0.20],[0.06,0.21,0.20],[0.18,0.22,0.20],[0.03,0.20,0.20]	35	[0.03,0.23,0.21],[0.46,0.20,0.20],[0.04,0.23,0.21],[0.46,0.26,0.21],[0.00,0.09,0.17]
6	[0.06,0.26,0.23],[0.86,0.23,0.22],[0.08,0.27,0.23],[0.00,0.23,0.22],[0.00,0.01,0.10]	36	[0.00,0.06,0.16],[0.88,0.22,0.21],[0.08,0.26,0.21],[0.00,0.22,0.21],[0.04,0.25,0.21]
7	[0.00,0.00,0.05],[0.88,0.23,0.24],[0.08,0.27,0.24],[0.00,0.24,0.23],[0.04,0.26,0.24]	37	[0.36,0.26,0.24],[0.00,0.00,0.03],[0.43,0.26,0.24],[0.00,0.23,0.24],[0.21,0.25,0.24]
8	[0.05,0.20,0.20],[0.86,0.21,0.20],[0.06,0.21,0.20],[0.00,0.18,0.20],[0.03,0.20,0.20]	38	[0.03,0.21,0.20],[0.41,0.18,0.20],[0.54,0.23,0.21],[0.00,0.18,0.20],[0.02,0.20,0.20]
9	[0.07,0.27,0.22],[0.89,0.23,0.22],[0.00,0.01,0.11],[0.00,0.24,0.22],[0.04,0.26,0.22]	39	[0.06,0.24,0.21],[0.83,0.21,0.21],[0.08,0.24,0.21],[0.00,0.08,0.17],[0.04,0.23,0.21]
10	[0.06,0.26,0.23],[0.83,0.22,0.23],[0.08,0.26,0.23],[0.00,0.00,0.07],[0.04,0.25,0.23]	40	[0.03,0.21,0.20],[0.42,0.18,0.20],[0.04,0.21,0.20],[0.51,0.24,0.21],[0.00,0.15,0.19]
11	[0.06,0.26,0.22],[0.86,0.22,0.22],[0.08,0.26,0.22],[0.00,0.23,0.21],[0.00,0.03,0.13]	41	[0.30,0.22,0.20],[0.28,0.17,0.20],[0.01,0.19,0.20],[0.39,0.23,0.20],[0.01,0.19,0.20]
12	[0.06,0.21,0.20],[0.83,0.18,0.20],[0.08,0.21,0.20],[0.00,0.19,0.20],[0.04,0.21,0.20]	42	[0.00,0.10,0.18],[0.88,0.21,0.21],[0.08,0.24,0.21],[0.00,0.21,0.20],[0.04,0.23,0.21]
13	[0.36,0.26,0.23],[0.00,0.00,0.10],[0.43,0.26,0.23],[0.00,0.23,0.22],[0.21,0.25,0.23]	43	[0.05,0.20,0.20],[0.62,0.17,0.20],[0.06,0.20,0.20],[0.25,0.22,0.20],[0.03,0.20,0.20]
14	[0.07,0.22,0.20],[0.88,0.19,0.20],[0.01,0.20,0.20],[0.00,0.19,0.20],[0.04,0.21,0.20]	44	[0.07,0.27,0.23],[0.89,0.23,0.23],[0.00,0.00,0.07],[0.00,0.24,0.23],[0.04,0.26,0.23]
15	[0.06,0.26,0.22],[0.83,0.22,0.22],[0.08,0.26,0.22],[0.00,0.01,0.12],[0.04,0.25,0.22]	45	[0.06,0.26,0.23],[0.83,0.22,0.23],[0.08,0.26,0.23],[0.00,0.00,0.06],[0.04,0.25,0.23]
16	[0.04,0.25,0.22],[0.51,0.21,0.22],[0.05,0.25,0.22],[0.40,0.28,0.23],[0.00,0.01,0.10]	46	[0.03,0.20,0.20],[0.40,0.17,0.20],[0.04,0.20,0.20],[0.32,0.22,0.20],[0.21,0.21,0.20]
17	[0.00,0.07,0.16],[0.88,0.21,0.21],[0.08,0.25,0.21],[0.00,0.22,0.21],[0.04,0.24,0.21]	47	[0.00,0.01,0.10],[0.88,0.23,0.22],[0.08,0.27,0.23],[0.00,0.24,0.22],[0.04,0.26,0.23]
18	[0.36,0.26,0.23],[0.00,0.00,0.09],[0.43,0.26,0.23],[0.00,0.23,0.22],[0.21,0.25,0.23]	48	[0.15,0.24,0.21],[0.58,0.09,0.18],[0.18,0.24,0.21],[0.00,0.21,0.20],[0.09,0.23,0.21]
19	[0.04,0.25,0.22],[0.58,0.22,0.22],[0.00,0.01,0.10],[0.35,0.28,0.23],[0.03,0.25,0.22]	49	[0.07,0.23,0.21],[0.89,0.20,0.20],[0.00,0.13,0.18],[0.00,0.21,0.20],[0.04,0.23,0.21]
20	[0.06,0.21,0.20],[0.76,0.18,0.20],[0.07,0.21,0.20],[0.08,0.21,0.20],[0.03,0.20,0.20]	50	[0.04,0.23,0.21],[0.49,0.20,0.20],[0.04,0.23,0.21],[0.00,0.09,0.17],[0.43,0.26,0.21]
21	[0.06,0.24,0.21],[0.86,0.21,0.21],[0.08,0.25,0.21],[0.00,0.22,0.21],[0.00,0.08,0.17]	51	[0.06,0.26,0.22],[0.86,0.22,0.21],[0.08,0.26,0.22],[0.00,0.23,0.21],[0.00,0.04,0.15]
22	[0.21,0.22,0.20],[0.31,0.17,0.20],[0.00,0.18,0.20],[0.47,0.23,0.21],[0.01,0.20,0.20]	52	[0.10,0.21,0.20],[0.80,0.18,0.20],[0.07,0.21,0.20],[0.00,0.19,0.20],[0.04,0.21,0.20]
23	[0.36,0.26,0.25],[0.00,0.00,0.00],[0.43,0.26,0.25],[0.00,0.23,0.25],[0.21,0.25,0.25]	53	[0.06,0.21,0.20],[0.83,0.18,0.20],[0.08,0.21,0.20],[0.00,0.19,0.20],[0.04,0.21,0.20]
24	[0.03,0.22,0.20],[0.34,0.19,0.20],[0.00,0.10,0.17],[0.34,0.25,0.21],[0.30,0.25,0.21]	54	[0.03,0.21,0.20],[0.33,0.18,0.20],[0.29,0.23,0.21],[0.00,0.15,0.19],[0.36,0.23,0.21]
25	[0.06,0.21,0.20],[0.76,0.18,0.20],[0.07,0.21,0.20],[0.08,0.21,0.20],[0.03,0.20,0.20]	55	[0.04,0.22,0.20],[0.51,0.19,0.20],[0.05,0.22,0.20],[0.40,0.24,0.21],[0.00,0.13,0.18]
26	[0.05,0.25,0.24],[0.69,0.22,0.23],[0.06,0.26,0.24],[0.19,0.27,0.24],[0.00,0.00,0.06]	56	[0.06,0.24,0.21],[0.86,0.21,0.21],[0.08,0.24,0.21],[0.00,0.21,0.20],[0.00,0.09,0.17]
27	[0.00,0.00,0.07],[0.88,0.23,0.23],[0.08,0.27,0.23],[0.00,0.24,0.23],[0.04,0.26,0.23]	57	[0.00,0.20,0.20],[0.88,0.19,0.20],[0.08,0.22,0.20],[0.00,0.19,0.20],[0.04,0.21,0.20]
28	[0.36,0.26,0.22],[0.00,0.00,0.11],[0.43,0.26,0.22],[0.00,0.23,0.22],[0.21,0.25,0.22]	58	[0.09,0.22,0.20],[0.72,0.13,0.19],[0.11,0.22,0.20],[0.02,0.21,0.20],[0.06,0.21,0.20]
29	[0.07,0.21,0.20],[0.87,0.18,0.20],[0.01,0.20,0.20],[0.01,0.20,0.20],[0.04,0.21,0.20]	59	[0.07,0.25,0.21],[0.89,0.21,0.21],[0.00,0.07,0.16],[0.00,0.22,0.21],[0.04,0.24,0.21]
30	[0.06,0.26,0.24],[0.83,0.22,0.24],[0.08,0.26,0.24],[0.00,0.00,0.05],[0.04,0.25,0.24]	60	[0.06,0.23,0.21],[0.83,0.20,0.20],[0.08,0.23,0.21],[0.00,0.13,0.18],[0.04,0.22,0.21]

(4) Iterative Update

Using the initial membership degrees, the four parameters (*Ex*, *En*, *He*, *β*) of each feature cloud set of the clustering centers were updated, yielding new clustering centers. For the feature value of each text object among the initial feature values in Table 1, the membership degree interval on the cloud set of the corresponding feature of the new clustering centers was calculated according to the calculation formula of upper and lower membership degrees. And their upper, medium and lower membership degrees were normalized respectively. Subsequently, the upper, medium and lower membership degrees of the normalized membership degrees of the five features of each text were multiplied respectively, yielding comprehensive membership degrees. Furthermore, the upper, medium and lower membership degrees of the comprehensive membership degrees were normalized by clustering centers, yielding the normalized results of the comprehensive membership degrees.

By comparing the previously calculated normalized comprehensive membership degrees with the currently calculated ones, determine whether the sum of the absolute differences between the corresponding upper, medium and lower membership degrees at the same positions of the two sets of results is less than the threshold of 0.0001. If so, the iteration is stopped. Depending on the specific requirements, the classification membership degrees and threshold settings may be determined using one or more of formulas (44–47). In this study, each text is assigned to a cluster center based on the criterion that the mean upper membership degree of its features is greater than 0.5. The minimum text count per class is set to 3. If a text satisfies the criteria for multiple cluster centers, the additional rule of selecting the center with the maximum sum of lower membership degrees is adopted.

Repeat the above iterative process until the iteration is stopped. Based on the finally obtained normalized comprehensive membership degrees, each text is assigned to the clustering center with the maximum mean of the upper and lower membership degrees in the results of normalized comprehensive membership degrees. After iterations, iteration differences of upper, medium, and lower membership between iteration *t* and *t*-1 are shown in Figure 2. The number of iterations is relatively small, owing to the appropriate selection of initial clouds and initial cluster centers.

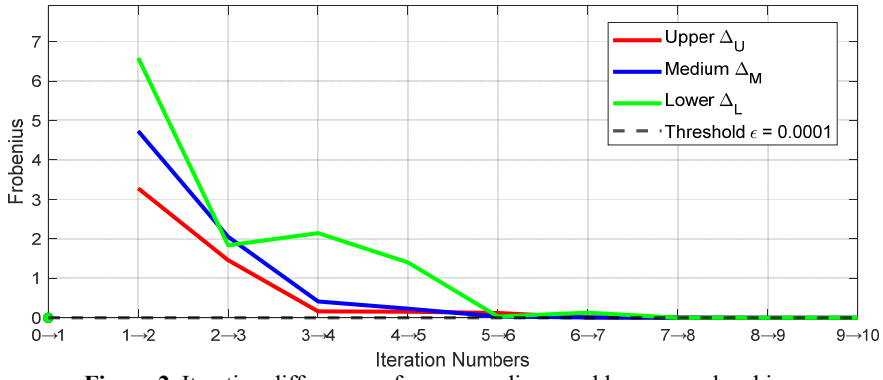


Figure 2. Iteration differences of upper, medium, and lower membership

The following five final clustering centers in Table 9 were obtained:

Table 9. Five final clustering centers

Cluster center	F1 (Ex,En,He, β)	F2(Ex,En,He, β)	F3(Ex,En,He, β)	F4(Ex,En,He, β)	F5(Ex,En,He, β)
CC1	C1-1(0.22,0.20,0.03,8.07)	C2-1(0.15,0.12,0.01,14.52)	C3-1(0.17,0.16,0.03,6.45)	C4-1(0.18,0.14,0.03,5.65)	C5-2(0.62,0.19,0.03,7.66)
CC2	C1-1(0.22,0.20,0.03,8.07)	C2-1(0.15,0.12,0.01,14.52)	C3-1(0.17,0.16,0.03,6.45)	C4-2(0.57,0.26,0.03,10.49)	C5-1(0.18,0.17,0.03,6.86)
CC3	C1-1(0.22,0.20,0.03,8.07)	C2-1(0.15,0.12,0.01,14.52)	C3-2(0.60,0.16,0.05,3.87)	C4-1(0.18,0.14,0.03,5.65)	C5-1(0.18,0.17,0.03,6.86)
CC4	C1-1(0.22,0.20,0.03,8.07)	C2-2(0.63,0.20,0.03,8.07)	C3-1(0.17,0.16,0.03,6.45)	C4-1(0.18,0.14,0.03,5.65)	C5-1(0.18,0.17,0.03,6.86)
CC5	C1-2(0.76,0.20,0.03,8.07)	C2-1(0.15,0.12,0.01,14.52)	C3-1(0.17,0.16,0.03,6.45)	C4-1(0.18,0.14,0.03,5.65)	C5-1(0.18,0.17,0.03,6.86)

Finally, the classification results of the texts were obtained as shown in Table 10 below:

Table 10. Clustering statistics

Cluster	Number of Texts	Text set
Hydrogen Policy	13	1,6,11,16,21,26,31,35,40,46,51,55,56
Hydrogen Infrastructure	11	5,10,15,20,25,30,34,39,45,50,60
Hydrogen Cooperation	7	9,14,33,38,44,49,59
Hydrogen Technology	11	3,8,13,18,23,28,37,43,48,53,58
Hydrogen Production	11	2,7,12,17,27,32,36,42,47,52,57
Outlier	7	4,19,22,24,29,41,54

The clustering method achieves a 93.33% overall recall rate when including outliers, correctly identifying 56 out of 60 samples after optimal matching via the proposed algorithm in Table 11. This demonstrates that the proposed approach effectively captures the underlying thematic structure of the hydrogen energy document corpus.

Table 11. Recall rate for standard normal cloud set C-mean clustering

Category	True Count	Correct Cluster Count	Recall Rate
Hydrogen Policy	12	12	100.00%
Hydrogen Infrastructure	12	11	91.67%
Hydrogen Cooperation	10	7	70.00%
Hydrogen Technology	11	11	100.00%
Hydrogen Production	11	11	100.00%
Outlier(no separate cluster)	4	4	100.00%
Total	60	56 (Including Outliers)	93.33%
		52 (Excluding Outliers)	86.67.%

5.2. Comparative Analysis

The traditional Fuzzy C-means (FCM) clustering method was applied to the same case for competitive intelligence clustering analysis, with the corresponding parameter settings as follows: the number of clusters *c*=5, the fuzziness coefficient *m*=2, the convergence threshold epsilon=1e-4, and the maximum number of iterations t= 1000. The fuzzy membership matrix *U* was initialized as follows: for 60 texts × 5 clusters, the membership degrees were randomly

generated, with the sum of each row equal to 1. The objective function of FCM clustering is $J^{(t)}=\sum_{i=1}^{n}\sum_{j=1}^{h} u_{i,k}{}^{m}\|x_i\text{-}v_k\|^2$. And the formulas $v_{k,j}{}^{(t)}=\frac{\sum_{i=1}^{n} u_{i,k}{}^{(t-1)^m} x_{i,j}}{\sum_{i=1}^{n} u_{i,k}{}^{(t-1)^m}}$ and $u_{i,k}{}^{t}=\frac{1}{\sum_{i=1}^{n}\frac{\|x_j\text{-}v_k\|^{\frac{2}{m-1}}}{\|x_i\text{-}v_l\|}}$ are adopted for the iterative update.

Single-sample clusters meeting the conditions are retained, with the minimum text count per class set to 1. The classification threshold of membership degree is set to 0.5. Texts that fall below the threshold and remain unassigned are classified as outliers. After 59 iterations, the clustering results are shown in Table 12 below:

Table 12. The results of clustering for FCM

Cluster	Number of Texts	Text set
Hydrogen Policy	11	1, 6, 11, 16, 21, 26, 35, 40, 51, 55, 56
Hydrogen Infrastructure	14	5, 15, 20, 25, 29, 31, 34, 39, 43, 46, 50, 54, 58, 60
Hydrogen Cooperation	6	9, 19, 33, 44, 49, 59
Hydrogen Technology	6	3, 13, 18, 23, 28, 37
Hydrogen Production	8	2, 7, 17, 27, 32, 36, 42, 47
Outlier (no separate cluster)	15	4,8,10,12,14,22,24,30,38,41,45,48,52,53,57

Feature values of cluster centers for FCM are shown in Table 13:

Table 13. Feature values of final cluster centers

Cluster center	F1	F2	F3	F4	F5
FCC1	0.0067	0.0029	0.0071	0.1014	0.5264
FCC2	0.0152	0.0240	0.0315	0.3027	0.0393
FCC3	0.0273	0.0052	0.4698	0.0584	0.0164
FCC4	0.0074	0.5179	0.0064	0.0171	0.0051
FCC5	0.7338	0.0020	0.0071	0.0091	0.0280

It can be seen from the convergence curve of the FCM objective function in Figure 3 that the objective function decreases monotonically with the growth of iteration times, which satisfies the theoretical property of FCM. The curve gradually flattens out after 59 iterations, and a more reasonable clustering partition is finally obtained.

It can be seen from Figure 4 that the F-norm difference gradually declines throughout iterations and finally falls below the threshold $\epsilon = 10^{-4}$ at the 59-th iteration.

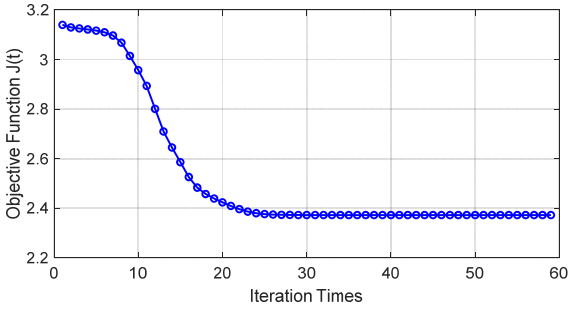


Figure 3. Convergence curve of FCM objective function

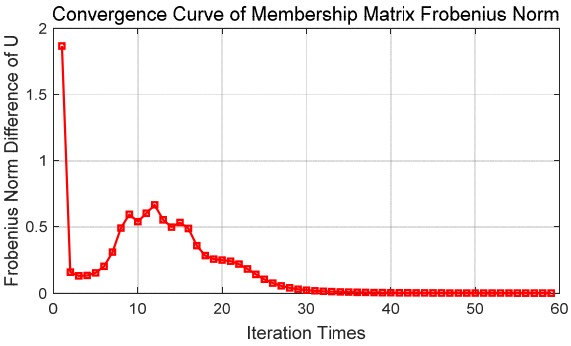


Figure 4. Convergence curve of membership matrix update difference

Recall rate for FCM clustering can be observed in following Table 14. Compared with the proposed method, FCM exhibits a significant performance gap of in terms of overall recall rate.

Table 14. Recall rate for FCM clustering

Category	True Count	Correct Cluster Count	Recall Rate
Hydrogen Policy	12	10	83.33%
Hydrogen Infrastructure	12	8	66.67%
Hydrogen Cooperation	10	6	60.00%
Hydrogen Technology	11	6	54.55%
Hydrogen Production	11	8	72.73%
Outlier	4	3	75.00%
Total	60	41 (Including Outliers)	68.33%
		38 (Excluding Outliers)	63.33%

The specific comparisons between cloud set and fuzzy set are elaborated in Table 15 as follows:

Table 15. Comparisons of clusters between cloud set and fuzzy set

Comparison Dimension	Randomness	Fuzziness	Classification Logic	Membership degree	Inter-cluster separation	Intra-cluster compactness	Classified samples/Outliers	Average texts per class
FCM	No	Yes	Forced clustering	Single	Common	Good	Poor	Poor
Cloud Set C-mean	Yes	Yes	Data-driven	Triple	Good	Good	Common	Common

6. Conclusion

In conclusion, this study has accomplished three core objectives: Firstly, it proposes a new strictly standardized normal cloud set, systematically defining its theoretical framework and successfully deriving the specific parameter values through rigorous mathematical deduction and validation. Secondly, it develops an improved C-means clustering method by integrating this strictly standardized normal cloud set, establishing a complete algorithmic process that combines the advantages of cloud set uncertainty processing and traditional clustering logic. Thirdly, it verifies the practical applicability of the proposed model and method through a concrete application case in hydrogen energy scientific and technological intelligence text analysis, demonstrating the entire process from text preprocessing to clustering result interpretation.

The research exhibits three prominent advantages. First, the strictly standardized normal cloud set effectively unifies the processing of randomness and fuzziness inherent in complex text data, overcoming the limitations of traditional methods that fail to simultaneously consider both types of uncertainty and thus improving the robustness of data representation. Second, the integrated C-means clustering method preserves the interpretability of cloud models while maintaining the efficiency of clustering algorithms, enabling more accurate classification of hydrogen energy intelligence texts and supporting the extraction of valuable thematic information. Third, the successful application to hydrogen energy intelligence analysis fills the research gap in applying cloud set theory to this domain, offering a new technical approach for industry-oriented information mining and decision support.

However, the study also has certain limitations. First, the

parameter solving process relies on specific experimental datasets. Its adaptability to wider types of text data (such as multilingual hydrogen energy literature or low-quality unstructured texts) remains to be further verified. Second, the current clustering method does not fully consider the dynamic evolution characteristics of hydrogen energy technology hotspots, resulting in reduced effectiveness in capturing time-sensitive thematic changes in intelligence texts. In addition, in application case, each text class is associated with a single feature word group. For future work, it is worth investigating the case where each text class is associated with multiple feature word groups. Future research can focus on optimizing the parameter adaptive mechanism, integrating temporal features, and improving algorithm efficiency to overcome these limitations and enhance the practical value of the proposed method.

Acknowledgments

This study is funded by Basic Special Project for Public Welfare-Oriented Scientific Research Institutes of Fujian Province (No.2024R1008002).

References

[1] Chakraborti, S. J., & Dey, S. O. (2016). Multi-level K-means text clustering technique for topic identification for competitor intelligence. In 10th IEEE International Conference on Research Challenges in Information Science (pp. 1-10).

[2] Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy C-means clustering algorithm. Computers & Geosciences, 10(2-3), 191-203.

[3] Krishnapuram, R., & Keller, J. M. (1993). A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems, 1(2), 98-110.

[4] Kim, E. H., Oh, S. K., & Pedrycz, W. (2018). Design of reinforced interval type-2 fuzzy C-means-based fuzzy classifier. IEEE Transactions on Fuzzy Systems, 26(5), 3054-3068.

[5] Zeng, W., Ma, R., Yin, Q., Zheng, X., & Xu, Z. (2020). Hesitant fuzzy C-means algorithm and its application in image segmentation. Journal of Intelligent & Fuzzy Systems, 38(5), 5681-5692.

[6] Wu, C. M., & Guo, X. K. (2021). A novel single fuzzifier interval type-2 fuzzy C-means clustering with local information for land-cover segmentation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 5903-5917.

[7] Xu, J. D., Zhao, T. Y., Feng, G. Z., Ni, M. Y., & Ou, S. F. (2021). A fuzzy C-means clustering algorithm based on spatial context model for image segmentation. International Journal of Fuzzy Systems, 23, 816-832.

[8] Li, L. Q., Liu, Z. H., & Guo, K. (2016). Fuzzy C-means algorithm based on grey relational analysis. Journal of Fuzhou University (Natural Science Edition), 44(2), 170-175.

[9] Liu, H., & Zeng, P. F. (2021). A soft-sensing method for endpoint carbon content and temperature in converter steelmaking based on WGRA-FCM sample similarity measurement. Control and Decision, 36(9), 2170-2178.

[10] Li, M. N. (2025). A grey principal component clustering model based on trend improvement. Natural Science, 13(2), 231-242.

[11] Qin, K., Xu, M., & Li, D. (2007). Cloud model based fuzzy C-means clustering and its application. In Proceedings of SPIE-Geoinformatics 2007: Remotely Sensed Data and Information (Vol. 6752, Article 67520T).

[12] Yang, X., Xie, W. X., & Huang, J. J. (2008). A C-means clustering approach based on cloud model. In 2008 IEEE International Conference on Fuzzy Systems (pp. 965-968).

[13] Song, Y. Y., Li, C. B., & Qi, Z. Q. (2014). Extraction of power load patterns based on cloud model and fuzzy clustering. Power System Technology, 38(12), 3378-3383.

[14] Liu, Y., Liu, Z., Li, S., Guo, Y., Liu, Q., & Wang, G. (2023). Cloud-Cluster: An uncertainty clustering algorithm based on cloud model. Knowledge-Based Systems, 263, 110261.

[15] Mao, W. W., & Xu, K. J. (2024). Enhancement of the classification performance of fuzzy C-means through uncertainty reduction with cloud model interpolation. Mathematics, 12(7), 975. https://doi.org/10. 3390/ math 120 70975

[16] Anjali, A., & Gupta, A. (2025). A novel Shannon entropy-based backward cloud model and cloud K-means clustering. The Journal of Supercomputing, 81(1), 65.

[17] Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning—III. Information Sciences, 9(1), 43-80.

[18] Torres-Blanc, C., Hernandez-Varela, P., & Cubillo, S. (2018). Self-contradiction for type-2 fuzzy sets whose membership degrees are normal and convex functions. Fuzzy Sets and Systems, 352, 73-91.

[19] D'Alterio, P., Garibaldi, J. M., John, R. I., & Pourabdollah, A. (2021). Constrained interval type-2 fuzzy sets. IEEE Transactions on Fuzzy Systems, 29(5), 1212-1225.

[20] Hernandez, P., Cubillo, S., & Torres-Blanc, C. (2022). A complementary study on general interval type-2 fuzzy sets. IEEE Transactions on Fuzzy Systems, 30(11), 5034-5043.

[21] Wang, L. X. (2017). A new look at type-2 fuzzy sets and type-2 fuzzy logic systems. IEEE Transactions on Fuzzy Systems, 25(3), 693-706.

[22] Khairuddin, S. H., Hasan, M. H., Akhir, E. A. P., & Hashmani, M. A. (2022). Generating type-2 trapezoidal fuzzy membership function using genetic tuning. CMC-Computers, Materials & Continua, 71(1), 717-734.

[23] Mandal, S., Karim, I. U., & Raha, S. A. (2021). Theory of approximate reasoning with type-2 fuzzy set. Journal of the Indonesian Mathematical Society, 27(1), 9-28.

[24] Al-Mallah, M., Ali, M., & Al-Khawaldeh, M. (2022). Obstacles avoidance for mobile robot using type-2 fuzzy logic controller. Robotics, 11(6), 130.

[25] Abiyev, R. H., Sadikoglu, G., Alsalihi, A., & Abizada, R. (2022). Sensory evaluation of customer satisfaction using type-2 fuzzy logic. Journal of Intelligent & Fuzzy Systems, 43(4), 4541-4554.

[26] Baskov, O. V., & Noghin, V. D. (2022). Type-2 fuzzy sets and their application in decision-making: Implementations. Scientific and Technical Information Processing, 49(5), 292-300.

[27] Zarinbal, M., & Fazel Zarandi, M. H. (2014). Type-2 fuzzy image enhancement: Fuzzy rule based approach. Journal of Intelligent & Fuzzy Systems, 26(5), 2291-2301.

[28] Li, D. Y., Meng, H. J., & Shi, X. M. (1995). Subordinate cloud and subordinate cloud generator. Journal of Computer Research and Development, (6), 15-20.

[29] Li, D. Y., & Liu, C. Y. (2004). On the universality of the normal cloud model. Engineering Sciences, 6(8), 28-34.

[30] Li, D. Y., Liu, C., & Gan, W. (2009). A new cognitive model: Cloud model. International Journal of Intelligent Systems, 24(3), 357-375.

[31] Di, K. C., Li, D. Y., & Li, D. R. (1999). Cloud theory and its application in spatial data mining and knowledge discovery. Journal of Image and Graphics, 4(11), 32-37.

[32] Yue, X., Sun, Z. L., Zhang, Y. Q., & Li, G. Z. (2001). Web log data mining technology based on cloud model. Application Research of Computers, (11), 113-116.

[33] Wu, A. Y., Ma, Z. G., & Zeng, G. P. (2016). Set pair fuzzy decision method based on cloud model. Chinese Journal of Electronics, 25(2), 215-219.

[34] Gao, H. B., Zhang, X. Y., Liu, Y. C., & Li, D. Y. (2017). Longitudinal control for mengshi autonomous vehicle via Gauss cloud model. Sustainability, 9(12), 2259.

[35] Liu, H. C., Luan, X., Li, Z. W., & Wu, J. N. (2018). Linguistic Petri nets based on cloud model theory for knowledge representation and reasoning. IEEE Transactions on Knowledge and Data Engineering, 30(4), 717-728.

[36] Wang, H. L. (2020). Research on a cloud model set theory method referring to fuzzy set. Journal of Intelligent Systems, 15(3), 507-513.

[37] Wang, H. L., Fei, L. G., & Feng, Y. Q. (2022). Research on cloud relationship matrix and cloud relationship formula. Fuzzy Systems and Mathematics, 36(2), 36-48.

[38] Wang, H. L., Feng, Y. Q., & Fei, L. G. (2024). Uncertain linear programming with cloud set constraints integrating fuzziness and randomness. Applied Soft Computing, 151, 111157.

[39] Wang, H. L., Fei, L. G., & Feng, Y. Q. (2024). Intelligent control decision integrating fuzziness and randomness for automatic management of cash flow. PLOS ONE, 9(3), e0292748.

[40] Wang, H. L. (2027). Uncertain linear regression with cloud set coefficients with random membership degree. Journal of Computational and Applied Mathematics, 489, 117764.

[41] Luo, Z. Q., & Zhang, G. W. (2007). A new algorithm of backward normal one-variate cloud. Journal of Frontiers of Computer Science and Technology, 1(2), 234-240.

[42] Xu, C. L., Wang, G. Y., & Zhang, Q. H. (2014). A new multi-step backward cloud transformation algorithm based on normal cloud model. Fundamenta Informaticae, 133(1), 55-85.