IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 17 February 2024, accepted 4 March 2024, date of publication 7 March 2024, date of current version 26 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3374770

RESEARCH ARTICLE

Enhancing Web Text Clustering Accuracy and Efficiency With a Maximum Entropy Function Model: Overcoming High-Dimensional and Directional Challenges

XUMIN ZHAO[1,2,3], GUOJIE XIE[1], YI LUO[1,2], FENGHUA LIU[4], AND HONGPENG BAI[5]
[1]Key Laboratory of Open Data of Zhejiang Province, Hangzhou 310000, China
[2]College of International Business, Zhejiang Yuexiu University, Shaoxing 312000, China
[3]College of International Business, Philippine Christian University, Manila 0900, Philippines
[4]Huzhou Vocational and Technical College, Huzhou 313000, China
[5]School of Intelligence and Computing, Tianjin University, Tianjin 300072, China

Corresponding author: Guojie Xie (xieguojie1698@dingtalk.com)

This work was supported in part by Huzhou Science and Technology Plan Project (Research and Application of Multi-Level Integration and Sharing of Big Data in the Elderly Care Industry Chain Under Multiple and Complex Scenarios) under Grant 2022GZ57, and in part by the Science and Technology Research Program of Huzhou Vocational and Technical College under Grant 2022GY06.

ABSTRACT With the rapid development of large models such as Chatgpt, text clustering has become an important research topic in data mining. However, traditional clustering algorithms face challenges in terms of text clustering due to the high dimensionality and directionality of text data; in particular,the research on web text mining is insufficient,so the accuracy and efficiency of clustering algorithms need to be improved. Aiming at the above challenges,this paper proposes a maximum entropy function model and applies it to web text clustering to overcome these challenges and achieve better clustering results. Unlike the traditional clustering algorithm,this algorithm avoids the local minimum and realizes the global minimum. This study will help strengthen web text mining and provide valuable insights for future research. In summary,this paper proposes a novel text clustering method,MEMC, which uses the maximum entropy function model to overcome the challenges of high-dimensional and directional features. Compared with the popular algorithms in the international standard datasets,the method is 15% higher than the current popular k-means algorithm in purity and 6% higher than the AP algorithm.

INDEX TERMS Maximum entropy, mean clustering algorithm, high-dimensional data, directional features, neighborhood message propagation.

I. INTRODUCTION

The ever-expanding volume of data presents an immense challenge in the modern era, calling for effective management of this abundance of information. In the realm of exploratory data analysis [1], [2], clustering emerges as a valuable tool across various domains, encompassing pattern recognition [3], feature extraction [4], vector quantization (VQ) [5], image segmentation [6], function approximation [7]. Data mining [8], [9]. In the context of addressing the challenges posed by big data, large-scale data clustering assumes a central role [10].

Spatial data clustering techniques have been widely explored to uncover meaningful patterns from intricate real-world data sources [11]. Notably, the DBSCAN algorithm, a renowned density-based clustering approach, holds a pioneering status in this field, requiring only a solitary input parameter, thereby granting users the flexibility to determine suitable values [12]. This remarkable characteristic enhances the accuracy and efficiency of clustering by accommodating

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

data with arbitrary shapes while effectively identifying noise samples within potential datasets [13]. However, the algorithm's high time complexity poses challenges when dealing with large and high-dimensional databases.

Since Li et al. [14] studied entropy-based classification data clustering criteria in 2004, entropy-based text implementation is not understood, and more and more researchers have paid attention to it. Carretero et al. used the Shannon information entropy to quantify the information content in the order each word appears in a text. Singhal et al. cite singhal2021keyword domain independent keyword extraction method based on rsamnyi entropy. The proposed word ranking metric's actual performance and relative performance are discussed. 2023 Giri and Majumder [15] explored the scope of application of feature extraction and maximum entropy-based fuzzy clustering (MEFC) in eigenvalue-based collaborative spectrum sensing (CSS).

In summary, existing challenges in web text clustering include the high dimensionality and directionality of web text data, data sparsity, scalability, subjective determinations, human-related challenges, and difficulty processing multilingual or mixed-language content. Addressing these challenges is essential to improving the effectiveness and efficiency of web text clustering algorithms.

Consequently, this study aims to address this limitation by introducing the OP-DBSCAN algorithm, which segregates the dataset into an operation set and potential datasets, thereby mitigating the runtime of the DBSCAN algorithm. The proposed approach leverages local data to identify neighboring elements and performs cluster identification steps utilizing smaller datasets. Considering these challenges, this research presents a novel approach for web text clustering, utilizing the Maximum Entropy Function Model (MEMC). By effectively tackling the hurdles posed by high-dimensionality and directional features inherent in web text data, this innovative methodology provides valuable insights for future investigations. It enhances our understanding of optimal solutions for web text clustering.

The specific innovations of this study can be summarized as follows:(1)Methodological Innovation: This study pioneers a groundbreaking web text clustering technique by harnessing the power of the Maximum Entropy Function Model (MEMC). By effectively addressing the challenges associated with high-dimensional and directional features in web text data, this methodology represents a profound advancement in traditional clustering algorithms.(2)Superior Accuracy and Efficiency: The proposed MEMC model exhibits remarkable improvements in clustering accuracy and efficiency, surpassing established algorithms such as k-means and AP by 15% and 6% respectively, in terms of purity. This assertion is supported by comprehensive evaluations conducted on internationally recognized datasets.(3)Achievement of Global Optimization: Diverging from conventional clustering algorithms that often fall into local minima, the proposed method leverages the maximum entropy function to attain global optimization. As a result, it produces more precise and meaningful clustering outcomes for web text data.

The paper's structure unfolds as follows: Section I provides a concise introduction to the background and significance of the study. In Section II,we outline an overview of the relevant literature, delineate the study's purpose and scope. Section III delves into the theoretical underpinnings of text clustering and traditional clustering algorithms. We expound on the maximum entropy function model and its application in web text clustering. Subsequently,Section IV presents a novel web text clustering method,explores its implementation and optimization,and presents and discusses experimental results. In Section V,we deliberate upon the outcomes,outline future research directions,and conclude the paper with acknowledgments and references.

II. RELATED WORK

With the exponential proliferation of text-based information sources,particularly with the advent of the internet,text mining has garnered significant attention from the academic community. Classification and clustering of text data,as integral components of text mining,hold particular importance. Textual information possesses vast storage capacity and undergoes rapid changes,making knowledge extraction a formidable task. Consequently,text mining has emerged as a prominent research area.

The rise of the internet has made electronic text storage an indispensable part of modern life. While advancements in storage devices continue,challenges in managing large-scale text information persist. Security concerns,such as the risk of hackers and system crashes,remain prevalent despite existing encryption and backup mechanisms. Jagtap and Ramudu proposed a secure storage scheme based on cryptography and neural networks [16],but it requires additional resources and power. Efficient retrieval of text data is also critical, and Oliver et al. suggested a retrieval method that utilizes document vector models and minimized hash to save time and resources [17]. Burkard et al. developed a technique to identify and prerender high-ranking search results; however, further optimization is necessary. Long-term preservation of electronic text proves to be another pressing issue,as storage cycles are contingent upon devices and technology, limiting long-term storage feasibility. Tan et al. proposed a promising DNA-based text preservation scheme [18] that holds potential in terms of long-term storage and scalability, despite the challenges at hand.

To summarize,electronic text storage plays a crucial role today,but challenges related to security, retrieval efficiency, and long-term preservation must be addressed.

A. LIMITATIONS OF TRADITIONAL TEXT CLUSTERING ALGORITHMS

Introduction: Text clustering is important in data mining,data science,and natural language processing [19].

Traditional clustering algorithms face numerous challenges when handling high-dimensional and directional text data. These challenges assume even greater significance in the realm of web text mining,which remains a relatively unexplored research area [20]. Consequently,there is a need to enhance the accuracy and efficiency of text clustering algorithms.

Traditional Text Clustering Algorithms: Various methods exist for text clustering,including algorithms such as K-means [21],hierarchical clustering [22],DBSCAN [23],and expectation maximization (EM) [24]. However,these traditional algorithms encounter limitations when applied to high-dimensional and directional data.

For instance,K-means is a widely adopted traditional clustering algorithm that assigns data points to the nearest centroid repeatedly. However,it is susceptible to getting trapped in local optima. Hierarchical clustering,on the other hand,constructs a tree-like structure of nested clusters and is suitable for small datasets. However,it lacks efficiency when confronted with large datasets. DBSCAN,another popular clustering algorithm,excels in identifying clusters with irregular shapes but requires pre-determined parameters,such as the minimum number of points to define a cluster,which can be challenging to ascertain in advance.

Related Work: To overcome the limitations of traditional clustering algorithms,several studies have proposed innovative algorithms for clustering high-dimensional and directional data. For example, Karim introduced a novel KNN-based approach for clustering high-dimensional data [25]. Their efficient algorithm eliminates local optima and converges to the global optimum,outperforming traditional K-means algorithms.

Another study by Li et al. presented two modified fuzzy clustering algorithms based on nonnegative matrix factorization,namely MFCM-NMF and MFCM-LCNMF [26]. These algorithms demonstrate enhanced clustering performance compared to traditional approaches. Additionally, Revanna et al. proposed an optimal data clustering method that combines particle swarm optimization with the JAYA approach,incorporating the concept of K-means clustering to initiate the search for optimal clusters [27].

In conclusion,text clustering is a vital research area,and traditional clustering algorithms face several challenges when confronted with high-dimensional and directional data. These limitations include high dimensionality,directionality,local optima,and sensitivity to noise. To address these challenges,researchers have proposed various novel algorithms, including the KNN-based approach,NMF-based algorithms,and fuzzy K-means algorithm. These innovative methods exhibit improved clustering performance compared to traditional algorithms. Future research efforts should focus on designing efficient and robust algorithms further to enhance the accuracy and efficiency of text clustering. Table 1 provides a comprehensive comparison of the effectiveness of various algorithms.

Therefore,future research should concentrate on the development of efficient and robust algorithms to enhance both the accuracy and efficiency of text clustering.

B. PRINCIPLE AND APPLICATION OF THE MAXIMUM ENTROPY FUNCTION MODEL

The maximum entropy function model is a probabilistic model that adheres to the maximum entropy principle,which provides a general method for selecting probability distributions [37], [38]. The principle has been further developed, demonstrating that the maximum entropy distribution minimizes the Kullback-Leibler divergence to achieve an even distribution. It takes on an exponential form, and its maximization is convex [39]. While inferring the complete distribution can be computationally hard,it can be approximated [40]. Exponential families and basis function expansions under moment constraints have shown promise for such approximations [41]. The counting problem can also be approximated to approximate the distribution. Additionally,relaxed torque constraints have been introduced through a maximum entropy problem formulation with generalized regularized measures in dual form [42]. There may be generalized constraints on noise as well. A strategy utilizing the duality of the maximum entropy problem and employing fast gradient approximation has been proposed [43]. Efficient inference has been explored through dynamically factorizing joint distributions,leading to accurate classification [44]. This expandable method can approach the original distribution and derive a simple pattern set,albeit with increasing computational complexity. However,specific experimental results and performance evaluations are lacking,hindering a comprehensive understanding of the method's performance across different datasets.

However,specific experimental results and data analyses were not explicitly mentioned, emphasising the method and theoretical foundations. Empirical research support is lacking,and potential limitations or issues were not identified.While the maximum entropy function model has been widely utilized and has demonstrated progress,there is ongoing research on improving its training algorithms and expanding its applications across different fields. Due to its straightforward principle and broad range of applications,the model still holds great research potential.

III. MODEL CONSTRUCTION

A. MAXIMUM ENTROPY MODEL

The maximum entropy model is a common classification algorithm in pattern recognition and statistical evaluation. It is a statistical model that follows the principle of maximum entropy. When predicting the probability distribution of a random event,the prediction should satisfy all known constraints and avoid making subjective assumptions about the unknown. In this context,the predicted probability distribution is the most uniform,resulting in the lowest predicted risk and the

TABLE 1. Comparison of clustering algorithms.

clustering algorithm	core thoughts	Method complementarity	Method defect	Representative work
K-means	By iterating to find k cluster centers, the distance between each data point and the nearest cluster center is minimized	Simple,fast and easy to implement; It has good effect on spherical distribution data	Sensitive to the selection of initial clustering centers; Inability to process non-spherically distributed data; Not suitable for dealing with noise and outliers	Image segmentation [28], customer clustering [29], gene classification [30],etc
DBSCAN	By finding density-connected data points to form clusters,clusters of arbitrary shapes can be found	Able to find clusters of arbitrary shapes; Not sensitive to noisy data	Sensitive to parameter selection; It may be less efficient for high-dimensional data and large-scale data	Anomaly detection [31], image segmentation [32], social network analysis [33], etc
HIDDEN	The proposed approach simultaneously learns classifier parameters and label embeddings, leading to improved performance over baseline methods. The learned hyperbolic embeddings accurately represent the label hierarchy,and the proposed classifiers achieve state-of-the-art generalization on standard benchmarks	It addresses the problem of label hierarchy without assuming prior knowledge of the hierarchy. This means that it can effectively handle complex label structures that may not be explicitly defined or known.	The joint learning approach simultaneously learns classifier parameters and label embeddings,leveraging the prior knowledge of label hierarchy and capturing the manifold structure of input data.	Information retrieval [34],Text mining and data analysis [35], [36]
HiLAP	ThThe method formulates HTC as a Markov decision process and learns a Label Assignment Policy via deep reinforcement learning to determine where to place an object and when to stop the assignment process. HiLAP makes inter-dependent decisions and can incorporate different neural encoders as base models for end-to-end training.	HiLAP incorporates deep reinforcement learning to learn a Label Assignment Policy,enabling inter-dependent decisions on where to place objects and when to stop the assignment process. The proposed method explores the hierarchy consistently during training and inference,addressing the mismatch between training and inference in existing HTC methods.	Computational cost: DRL is computationally expensive,especially when dealing with large-scale datasets.	Information retrieval [34],Text mining and data analysis [35], [36]
SOM	The model uses a feature matrix and a correlation matrix to explore the crucial dependencies between labels and generate classifiers for the task. Attention allows the system to assign different weights to neighbor nodes per label,enabling it to learn the dependencies among labels implicitly	Captures the attentive dependency structure among labels,addressing the issue of label dependencies that are often ignored by existing methods.	Difficulty with large datasets and labels: When the dataset contains a large number of labels,the correlation matrix used in the algorithm becomes oversized,making training challenging.	Information retrieval [34],Text mining and data analysis [35], [36]

TABLE 2. Main symbol table.

Symbol	Implication
$f(x,y)$	Feature Function
$\widetilde{\mathcal{R}}$	Training Data
$E_{\tilde{\mathcal{R}}}(f)$	Mathematical Expectation
T	Empirical Distribution Hypothesis
$\tilde{\mathcal{R}}(x)$	$\mathcal{R}(x)$ is approximate procedure
C_i	Constraint Condition
$\mathcal{J}_T(U,V)$	Cluster optimization
$\mathcal{H}(u)$	Entropy Function
$\mathcal{T}$	Annealing Coefficient
$\mathcal{J}_T$	Free Energy Function Of The System
X	Sample Set
$\mathcal{J}$	Global Maximization Cost Function
$\mathcal{U}_T(u,v)$	Minimization Objective Function

highest entropy. Below is Table 2,which lists the primary symbols used in this text for better comprehension:

The principle of maximum entropy prioritizes satisfying all known constraints in the model. The idea behind obtaining the reduced bundle is as follows: Several features are extracted from the training data,and the expectation of these features (given the feature function $f(x, y)$) on the training data regarding the empirical distribution (the distribution obtained statistically from the training data) (i.e.,the feature function $f(x, y)$ on the training data regarding $\widetilde{\mathcal{R}}$ (x,The mathematical expectation of $E_{\tilde{\mathcal{R}}}(f)$) of y and their expectation of $\mathcal{R}(x, y)$ in the model (i.e. the characteristic function $f(x, y)$) in the model concerning $\mathcal{R}$ (x,The mathematical expectation of y) is equal to $(E_{\mathcal{R}}(f))$. One feature corresponds to one constraint.

Empirical distribution hypothesis of data sets $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$,then The empirical distribution $\tilde{\mathcal{R}}(x, y)$ of the joint distribution $\mathcal{R}(x, y)$ is:

$$\tilde{\mathcal{R}}(x, y) = \tilde{\mathcal{R}}(X = x, Y = y) = \frac{\text{count}(x, y)}{N} \tag{1}$$

The empirical distribution of $\tilde{\mathcal{R}}(x)$ for the edge distribution of $\mathcal{R}(x)$ is approximate procedure used:

$$\tilde{\mathcal{R}}(x) = \tilde{\mathcal{R}}(X = x) = \frac{\text{count}(x)}{N} \tag{2}$$

Define the feature function $f(x, y)$ as:

$$f(x, y) = \begin{cases} 1 & x, y\text{satisfies a fact} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Then $\mathcal{E}_{\tilde{\mathcal{R}}}(f)$ is:

$$\mathcal{E}_{\tilde{\mathcal{R}}}(f) = \sum_{x,y} \tilde{\mathcal{R}}(x,y)f(x,y) \tag{4}$$

$\mathcal{E}_{\mathcal{R}}(f)$:

$$\begin{aligned}\mathcal{E}_{\mathcal{R}}(f) &= \sum_{x,y} \mathcal{R}(x,y)f(x,y) \\ &= \sum_{x,y} \mathcal{R}(x)\mathcal{R}(y \mid x)f(x,y) \\ &\approx \sum_{x,y} \tilde{\mathcal{R}}(x)\mathcal{R}(y \mid x)f(x,y)\end{aligned} \tag{5}$$

Then the constraints are proposed:

$$\begin{gathered}\mathcal{E}_{\tilde{\mathcal{R}}}(f) = \mathcal{E}_{\mathcal{R}}(f) \\ \sum_{x,y} \tilde{\mathcal{R}}(x,y)f(x,y) = \sum_{x,y} \tilde{\mathcal{R}}(x)\mathcal{R}(y \mid x)f(x,y)\end{gathered} \tag{6}$$

That is,the constraint condition C_i is (If n is extracted from the feature function,there are n feature function and n constraint).

$$C_i : \mathcal{E}_{\tilde{\mathcal{R}}}(f_i) = E_{\mathcal{R}}(f_i)\, i = 1, 2, \cdots, n \tag{7}$$

Another constraint is (input samples always belong to a certain class):

$$\sum_{y} \mathcal{R}(y \mid x) = 1 \tag{8}$$

Derivation of entropy (conditional entropy)

$$\begin{aligned}\mathcal{H}(\mathcal{R}) &= \sum_{x \in X} \mathcal{R}(x)\mathcal{H}(Y \mid X = x) \\ &= -\sum_{x \in X} \mathcal{R}(x) \sum_{y \in Y} \mathcal{R}(y \mid x) \log \mathcal{R}(y \mid x) \\ &= -\sum_{x,y} \mathcal{R}(x)\mathcal{R}(y \mid x) \log \mathcal{R}(y \mid x) \\ &\approx -\sum_{x,y} \tilde{\mathcal{R}}(x)\mathcal{R}(y \mid x) \log \mathcal{R}(y \mid x)\end{aligned} \tag{9}$$

The maximum entropy clustering algorithm (MEC) is a prominent example of incorporating entropy methods. Various versions of MEC may have different descriptions, but these differences are merely superficial.

For the data set, X is a membership degree matrix. u_{ij} is the probability that each sample belongs to the class center and satisfies:

$$\begin{gathered}u_{ij} \in [0, 1], \quad 1 \leqslant \mathbf{i} \leqslant \mathrm{K}, \quad 1 \leqslant \mathbf{j} \leqslant \mathbf{n}, \\ \sum_{i=1}^{K} u_{ij} = 1 \ldots\ldots\ldots\ldots\end{gathered} \tag{10}$$

The maximum entropy fuzzy clustering algorithm MEC divides N vector $x_i(i = 1, 2, \ldots, N)$ into K clusters.$G_i(i - 1, 2, \ldots, k)$,and the clustering center of each cluster is obtained,and the following objective functions are minimized:

$$\mathcal{J}_T(U,V) = \sum_{i=1}^{K}\sum_{j=1}^{N} u_{ij} \left\| x_j - v_i \right\|^2 + T \sum_{i=1}^{K}\sum_{j=1}^{N} u_{ij} \ln u_{ij} \tag{11}$$

The $left\ |x_j - v_i\ right\ |^2 = left(x_j - v_i\ right)^T\ left(x_j - v_i\ right)$, $mathrmT$ is though laser multiplier. The above equation can also be expressed as

$$\mathcal{J}_T = \mathcal{J}_c(U,V) - \mathcal{T}\mathcal{H}(u) \tag{12}$$

where $\mathcal{J}_c(U,V) = sum_{I=1}^{K}$. If the clustering problem is regarded as a physical system,then $\mathcal{J}_c(U,V)$ is equivalent to the energy in the deterministic annealing technique,$\mathcal{H}(u)$ is the entropy function,and the Lagrange multiplier of this T is equivalent to the temperature coefficient of the deterministic annealing technique,also known as the annealing coefficient. $\mathcal{J}_T$ is the free energy function of the system. Obviously,for large T,the main attempt is to maximize the entropy $\mathcal{H}(u)$,the system is maintained at a high temperature,and the global minimum point of the system is easy to find. As T decreases, entropy is exchanged for distortion,and as T approaches zero,minimizing the energy function $\mathcal{J}_c(U,V)$ directly yields a non-random solution. The process is to solve the clustering problem by solving a series of minimum points of the free energy function which changes with temperature T.

The MEC algorithm can avoid the local minimum and get the global minimum,so it has been widely used. However,one of the defects of MEC algorithm is the use of European metric. For high-dimensional vectorized text data,the direction feature of the text vector is far more important than its size feature, so MEC is unsuitable for cluster analysis of text data.

B. MODEL CONSTRUCTION

Through the above analysis, we introduce the maximum entropy principle into the mean clustering algorithm to build the model of this paper: MEMC. The model construction process is as follows:

For the sample set $X = \{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^d$ and $x_i x_i^T = 1\ (1 \leq i \leq N)$, $V = \{V_1, V_2, \ldots, V_K\}$ is a cluster center of K and has $v_i^T v_i = 1\ (1 \leq i \leq N, 2 \leq K \leq N)$, $U = \{u_{ij}\}_{K \times N}$ is a membership matrix. u_{ij} is the membership degree of sample x_j belonging to the centre of K, and its value is different from the hard division of the mean value of K, but a fuzzy division between 0 and 1, thus reflecting the real relationship between data points and class centre points. And $\sum_{i=1}^{K} u_{ij} = 1$. In this case, the global maximization cost function can be regarded as:

$$\mathcal{J} = \sum_{i}^{k}\sum_{j}^{n} u_{ij} x_j^T v_i \tag{13}$$

To achieve the highest possible value of the formula above and avoid settling for a local minimum instead of the global minimum,the maximum entropy principle can be introduced. This involves defining the objective function for minimization.

$$\mathcal{J}_T(u,v) = -\sum_{i=1}^{K}\sum_{j=1}^{N} u_{ij} X_j^T V_i + \frac{1}{\mathcal{T}}\sum_{i=1}^{K}\sum_{j=1}^{N} u_{ij}\ln u_{ij} \quad (14)$$

Incorporating the entropy term in this formula highlights the utilization of the cosine aggregation effect quantity, which is more appropriate for text analysis than the limited Euclidean measure. Consequently, this formulation corresponds to the maximum entropy objective function, specifically tailored for text clustering. The representation of the formula as mentioned earlier can also be articulated as follows:

$$\mathcal{J}_T = \mathcal{J}_c(U,V) - \frac{1}{\mathcal{T}}\mathcal{H}(u) \quad (15)$$

where $\mathcal{J}_c(U,V) = -\sum_{i=1}^{K}\sum_{j=1}^{N} u_{ij}x_j v_i$, T is a Lagrange multiplier, which can be valued according to our needs, and its value has a certain influence on the final clustering result. $H(u)$ membership degree matrix of entropy,when $frac1T$ is large,minimize $\mathcal{J}_T(u,V)$actually need to maximize the entropy $H(u)$. As $frac1T$ value decreases,and minimize $\mathcal{J}_T(U,V)$ to minimize $\mathcal{J}_c(U,V)$,so as to obtain the global minimum point.

To find the minimum value of $\mathcal{J}_T(U,V)$ is actually to find the peak value of the objective function under the condition of $v_i^T v_i = 1(i = 1, 2, \ldots, K)$ and $\sum_i^K u_{ij} = 1$. To this end, the Lagrange multiplier λ is introduced. And define Lagrange $\mathcal{L}$(u, v, λ, γ) as follows:

$$\mathcal{L}(u,v,\lambda,\gamma) = \mathcal{J}_I(U,V) + \lambda\sum_{i=1}^{K}(v_i^T v_i - 1) + \gamma\sum_{j=1}^{N}\left(\sum_{i=1}^{K} u_{ij} - 1\right) \quad (16)$$

The partial derivative of each central vector v_i in $\mathcal{L}$(u, v, λ, γ) is:

$$\frac{\partial\mathcal{L}(u,v,\lambda,\lambda)}{\partial v_i} = -\sum_{j=1}^{N} u_{ij}x_j^T + 2\lambda v_i \quad (17)$$

Equation (16) shows that if the expression is equal to zero, the vector v_i can be calculated as:

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}x_j}{2\lambda} \quad (18)$$

And from equation (17), it can be further derived that: $v_i^{\mathcal{T}} v_i = 1$,which can be further derived from equation (17): For the partial derivatives of each uij in the cost function $(u, v, ,)$, we have:

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}x_j}{\sqrt{\left(\sum_{j=1}^{N} u_{ij}x_j\right)^{\mathcal{T}}\left(\sum_{j=1}^{N} u_{ij}x_j\right)}} \quad (19)$$

For partial derivatives are:

$$\frac{\partial\mathcal{L}(u,v,\lambda,\gamma)}{\partial u_{ij}} = -x_j^{\mathcal{T}} v_i + \frac{1}{\mathcal{T}}\left(\ln u_{ij} + 1\right) + \gamma \quad (20)$$

If the above expression is equal to zero, then:

$$\ln\left(u_{ij}\right) = \mathcal{T}x_j^{\mathcal{T}} v_i - (\mathcal{T}\gamma + 1) \quad (21)$$

Since $\sum_i^K u_{ij} = 1$,we can further derive:

$$u_{ij} = \frac{e^{\mathcal{T}_x^{\mathcal{T}} v_i}}{\sum_{i=1}^{K} e^{\mathcal{T}_x^{\mathcal{T}} v_i}} \quad (22)$$

We iteratively find the minimum of equation (15) by iterating through equation (17) and equation (21). This process of finding the minimum object function is called the maximum entropy K mean clustering algorithm,which is equivalent to solving the clustering problem by solving a series of minimum points of the free energy function suitable for text clustering that change with temperature T. The following is the flow of the clustering algorithm based on maximum entropy K mean:

Algorithm 1 MEMC Algorithm

```
Input  : $K$ $(2 \leqslant K < N)$, $v^{(0)} = \left\{v_i^{(0)}, v_i^{(0)}, \ldots, v_K^{(0)}\right\}$,
         Fuzzy partition matrix
         $U = \{u_{ij}\}_{K\times N}$ , $l = 0$, $l$, Maximum number
         of iterations $M$, Annealing coefficient $\mathcal{T}$,
         maximum annealing coefficient Max$\mathcal{T}$
         threshold, number of iterations $\gamma = 0$
Output: The final clustering result cluster
for $\mathcal{T} \neq Max\mathcal{T}$ do
    $u_{ij} = \frac{e^{\mathcal{T}x_j^{\mathcal{T}} v_i}}{\sum_{i=1}^{K} e^{\mathcal{T}x_j^{\mathcal{T}} v_i}}$;
    Update:$u_{ij}^{(l+1)}$;
    if $\max_i^{l+1}\left\|v_i^{(l+1)} - v_i^l\right\| < \varepsilon$ then
        $\mathcal{T} = \mathcal{T} - \Delta\mathcal{T}$; $v_i = \frac{\sum_{j=1}^{N} u_{ij}x_j}{\sqrt{\left(\sum_{j=1}^{N} u_{ij}x_j\right)^T\left(\sum_{j=1}^{N} u_{ij}x_j\right)}}$
        Update:$v_i^{(l+1)}$
    else
        if $l > M$ then
            $\mathcal{T} = \mathcal{T} - \Delta\mathcal{T}$;
        end
    end
end
return Clustering result cluster
```

Algorithm 1 shows the whole process of algorithm optimization. To ascertain the algorithm's efficacy, we establish a limit of 50 iterations, denoted as the Maximum M, and ultimately compute the average. The annealing coefficient $\mathcal{T}$ is assigned a value of 0.2, while the maximum annealing coefficient $\mathcal{T}$ is set at 0.8. In the subsequent section, we conduct a comprehensive analysis to evaluate the algorithm's robustness.

IV. EXPERIMENTAL ANALYSIS

Text data is the most prevalent and extensively utilized form among the vast information resources. The information presented here takes the form of text. In the realm of information-oriented retrieval,content mining primarily involves extracting knowledge from unstructured and semi-structured documents, specifically text mining. The objective is to enhance the quality of search results and assist users in filtering out irrelevant information. While content mining predominantly focuses on text mining,there is a scarcity of research concerning multimedia data mining,such as images,pictures, videos,and audio. Research in these areas typically pertains to graphics and image processing,audio and recognition,and video analysis. Thus, emphasising in-depth text data mining research becomes particularly crucial,as it carries significant theoretical and practical value. This study approaches the topic from the perspective of text mining. In this paper,the term ''document'' refers to a fragment of text that includes the title and abstract content after undergoing page processing,excluding multimedia data.

A. EVALUATION CRITERIA

A proficient clustering method can generate clusters of high quality, characterized by strong intra-cluster cohesion and minimal inter-cluster overlap. Generally,there are two criteria commonly used for evaluating the quality of clusters: internal quality evaluation and external evaluation.

External evaluation employs known classification label datasets to assess the quality of clustering. This involves comparing the original label data with the output clusters. The desirable outcome of external evaluation is aggregating data points with different class labels into separate clusters,while data points with the same class labels are grouped. Commonly used external evaluation criteria include entropy,purity,and various other indicators.

Entropy measures the degree of class mixing within a cluster. To compute it,the class distribution of data within each cluster is first determined. Specifically,for cluster i,the probability that its members belong to class j is calculated.

$$p_{ij} = \frac{m_{ij}}{m_i} \tag{23}$$

where m_i represents the number of all objects in the cluster i,and m_{ij} is the number of objects of class j in the cluster i. Using the class distribution,use the standard common form:

$$e_i = -\sum_{\sum -4}^{K} p_{ij} \log_2 p_{ij} \tag{24}$$

Compute the average value for each cluster "i," with "k" representing the total number of classes. The collective average of the entire cluster set is determined by obtaining the weighted sum of the individual cluster averages,where the weight corresponds to the number of samples within each cluster.

$$e = \sum_{2}^{K} \frac{m_i}{m} e_i \tag{25}$$

where K is the number of clusters,and m is the sum of data points within the cluster

Purity: Another measure of containing a single class object within a cluster. The purity of the cluster i is $p_i = \max_j p_{ij}$,and the total purity of the cluster is:

$$\text{purity} \ = \sum_{\sum_y}^{K} \frac{m_i}{m} p_i \tag{26}$$

B. CONFUSION MATRIX

Confusion matrix,also known as an error matrix,is a standardized format for evaluating precision expressed in n rows and n columns. Various evaluation indices are included,such as overall accuracy, cartographic accuracy,user accuracy, etc. These measures of accuracy reflect the results of text classification from different angles. In text accuracy evaluation,a confusion matrix is commonly used to compare the classification results with the actual measured values and display the accuracy of the classification results. The confusion matrix is calculated by comparing the position and classification of each observed pixel with the corresponding position and classification in the classified image. The basic structure of the confusion matrix is as follows: True Positive (TP): The model correctly predicts that a sample in a positive category will be in a positive category.False Negative (FN): The model incorrectly predicts a negative class for a sample that is a positive class.False Positive (FP): The model incorrectly predicts a positive category for a sample that is a negative category. True Negative (TN): The model correctly predicts a negative category for a sample that is a negative category.

This study mainly employs the following two indices: Accuracy: The total percentage of correct predictions (both positive and negative).

$$\text{Accuracy} \ = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{27}$$

While the accuracy rate can provide an overall assessment of accuracy,it may not be a reliable indicator when dealing with unbalanced samples. In such cases,a high accuracy rate can be misleading and even invalid.

Recall rate,on the other hand,focuses on the proportion of correctly predicted positive instances out of all the actual positive instances. It measures the ability to identify positive samples in the original dataset correctly. A high recall rate implies a greater effort to detect every object that should be identified as positive,even if it results in more false checks.

$$\text{Recall} \ = \frac{TP}{TP + FN} \tag{28}$$

TABLE 3. The top 10 categories in reuter.

Class name	Training set	Test set
Grain	433	149
Corn	182	56
Wheat	212	71
Earn	2877	1087
Acquisitions	1650	719
Trade	369	118
Money-fx	538	179
Interest	347	131
Crude	389	189
Ship	197	89

C. TEST DATA SET

To thoroughly analyze and juxtapose against alternative research outcomes,as well as to diligently substantiate the efficacy of the proposed model,the ensuing pair of commonly employed datasets have been chosen to ascertain the model's effectiveness in the realm of text clustering:

The Reuters Corpus: This meticulously curated corpus comprises textual materials encompassing the domain of economics,alongside an array of meticulously organized document topics derived from the esteemed Reuters News Network. In this undertaking,one document is designated as the training set,with the remainder as the test set. It is crucial to note that the frequency distribution of document categories within the corpus is disparate,as the most sizable category entails a solitary document,whereas most categories encompass a relatively meager number of documents. For the experiment,sole consideration is attributed to documents annotated with a solitary category,thereby forging an exclusive association between a given document and a solitary category. The assortment of randomly selected documents,as incorporated within the experiment,embody an assemblage of unlabeled documents,whilst the remaining documents consigned to the training set are selectively labeled. Table 3 proffers a comprehensive depiction of the document count contained within each class's training and test sets.

The Newsgroups dataset is one of the internationally recognized benchmarks in text classification,text mining,and information retrieval research. This dataset encompasses approximately 20, 000 distinct documents derived from an array of newsgroups,each thoughtfully segregated into 20 discrete collections,each focusing on divergent subjects. Notably,certain newsgroups exhibit a remarkable degree of thematic resemblance. Originally compiled by Ken Lang in the year 1995,this reservoir of knowledge comprises precisely 19, 997 messages contributed by ardent Internet users through the Usenet platform. These informative missives are meticulously allocated across those as mentioned earlier 20 different newsgroups (commonly referred to as 20NG),warranting an equitable distribution of 1, 000 messages per group (with a single exception containing 997 messages). Each distinct newsgroup undertakes a unique text category, thereby facilitating comprehensive and stratified analysis.

D. CONTRAST ALGORITHM

K-means algorithm is a renowned and admired machine learning algorithm, one of the ten classic algorithms utilized in this field. K-means holds great utility in modern development trends due to its ease of comprehension,outstanding clustering effect,potent ability to cope with huge volumes of data processing,and low algorithmic complexity. The most commonly implemented partition clustering method is K-means clustering. In the process of K-means clustering analysis,the initial step involves determining the requisite number of classes,followed by the commencement of iterative steps until every observed value has been associated with the corresponding class. In this type of clustering,the square of the distance is used to characterize the difference within the class, to minimize it. Therefore,the intra-class difference for each class is the sum of the squares of the Euclidean distances between all pairs of observations in the class,divided by the number of observations in the class. Since the calculation utilizes an iterative algorithm,the results obtained from each K-means cluster may differ remarkably,even if the required number of classes remains the same. This disparity arises because the initial selection of observations is random, thereby engendering varied results for each cluster.

Affinity Propagation (AP) algorithm is a clustering learning method grounded on the metric of similarity. Its principal objective is to discover an optimal set of exemplars,and to identify the underlying relationship between these exemplars and the similar types of cluster samples. The nearest neighbor message propagation clustering method views each sample as a data node within the network of nodes,and considers them as the initial cluster representative center,while calculating the similarity matrix based on some relevant metric. In the clustering process,two types of evidence messages are incessantly relayed and updated between each data node until an optimal set of class representative centers and a corresponding class of clusters appear.

If the similarity metric utilized between each data point adopts the negative Euclidean distance grounded on the distance measure,then the goal of the AP clustering method is completely consistent with the classical K-means clustering algorithm. However,there remain prominent differences in the basic principles. During the clustering process,the AP method seeks the best class representative points grounded on the propagation and accumulation of two variations of evidential information,whereas the K-means algorithm updates the cluster centre by minimizing the total cost of the class center and the data within the class. At the outset of the AP method,all data points are seen as potential representative centres,and the number of clusters does not need to be specified a priori,thereby avoiding the clustering results being subjected to the setting of the initial class centers and the number of clusters. In contrast,the K-means algorithm randomly selects several initial centres at the outset,and the clustering results are easily affected by the initial center selection. Comparatively,the AP method displays remarkable efficiency in resolving non-Euclidean space issues (like when the similarity measure matrix is asymmetrical or does not conform to the triangle inequality criterion), while K-means and other clustering algorithms strictly demand adherence to the basic requisites of Euclidean distance space.

E. EXPERIMENTAL RESULTS AND ANALYSIS

To confirm the efficacy of the program,this research employed a $64-bit$ Windows 10 operating system,an Intel $i7-8700$ high-performance processor,16GB RAM,an Nvidia GTX 1050Ti graphics card,and a 160GB SSD hard drive. The device configuration was utilized to implement the algorithm via MATLAB 2019.

To verify the algorithm's effectiveness and evaluate the results objectively,the accuracy of the training set classification and the recall rate are usually mutually decreasing. Similarly,recall rates are often sacrificed for higher accuracy. It may be misleading to evaluate one of them individually. A more accurate and objective evaluation method is to take the F_1 index to consider both of them. After Reuter selected the top 10 categories,the small categories in the original corpus no longer dominated,while the size of each category in the Newsgroups dataset was relatively balanced,so most categories in the two data sets participating in the experiment were relatively balanced,and the micro-average index emphasized the impact of categories on the overall result. Therefore,using the micro-average F_1 index is more suitable for comprehensively reflecting the improved method's classification effect. The F_1 indicator of the micro average is defined as follows:

Objectively evaluate the results and verify the algorithm's efficiency, classification accuracy and recall rates are usually inversely related. As such,sacrificing recall rates is often necessary to achieve higher accuracy,and evaluating either metric individually may be misleading. A more accurate and objective evaluation method is considering both metrics using the F_1 index. After Reuter selected the top 10 categories,the smaller categories in the original corpus no longer dominated,and the size of each category in the Newsgroups dataset was relatively balanced. As a result,most categories in both datasets used in the experiment were relatively balanced,and the micro-average index emphasized the impact of categories on the overall outcome. Therefore,utilizing the micro-average F_1 index is more suitable for comprehensively reflecting the classification effect of the improved method. The F_1 score for the micro-average is defined as follows:

$$F_1^{\mu} = \frac{2 \times \mathrm{Pr}^{\mu} \times \mathrm{Re}^{\mu}}{\mathrm{Pr}^{\mu} + \mathrm{Re}^{\mu}} \tag{29}$$

where Pr^n is the micro-average accuracy and Re^u is the micro-average recall rate.

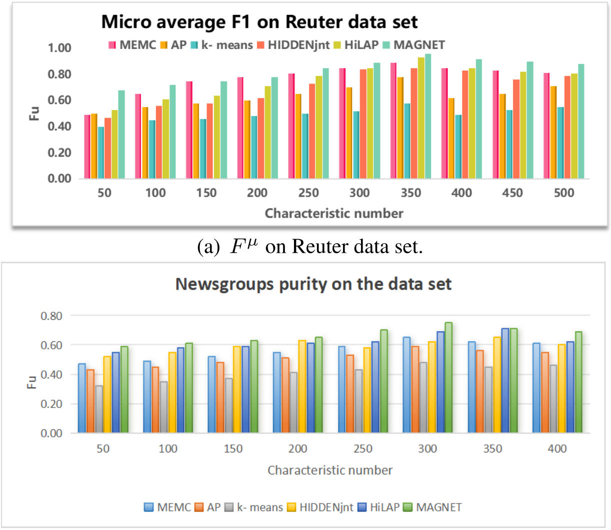
This study provides a comprehensive analysis of the clustering efficacy of the MEMC algorithm by comparing its performance with other popular clustering algorithms. The evaluation was conducted using two criteria: average index and clustering Figure 1(a) visualizes the integration of the MEMC algorithm with the maximum entropy principle, as well as the inclusion of other algorithms such as the traditional AP method, K-means clustering algorithm, HIDDENjnt, HiLAP, and MAGNET. These algorithms were applied to two distinct datasets to assess their strengths and weaknesses, and their performance was measured using the F_1 index.The results, as depicted in Figure 1(a), demonstrate the superiority of the MEMC clustering algorithm. Leveraging the maximum entropy principle effectively constructs a more efficient similarity measurement method, even with limited supervised prior information. The MEMC algorithm consistently achieves superior clustering outcomes across various types of datasets.In addition to its superior performance, the MEMC algorithm offers several advantages over the other algorithms. It successfully combines the strengths of traditional AP and K-means clustering algorithms with the innovative approaches of HIDDENjnt, HiLAP, and MAGNET. This integration provides a robust and versatile clustering solution that adapts well to different datasets. Furthermore, the MEMC algorithm's ability to work with a small amount of supervised prior information makes it particularly useful in real-world scenarios where acquiring extensive labeled data may be challenging or expensive. Its efficient similarity measurement method ensures accurate clustering results, which can have significant implications in various domains such as data analysis, pattern recognition, and recommendation systems.Overall, the results of this study emphasize the efficacy and versatility of the MEMC clustering algorithm. Its utilization of the maximum entropy principle and its superior performance across diverse datasets underlines its potential as a valuable tool for clustering tasks.

In this study, the clustering efficacy of the MEMC algorithm was evaluated through experimentation on two distinct datasets: Reuter and Newsgroups. These datasets possess different characteristics, allowing for a comprehensive algorithm performance analysis.The evaluation was based on the micro-average F index, which measures the algorithm's ability to classify instances across all categories correctly. The results revealed that the MEMC algorithm consistently outperformed the K-means and classical AP algorithms regarding the F index.Interestingly, as the number of strong category features increased, the MEMC algorithm reached its peak F index faster than the AP and K-means algorithms. This finding highlights the exceptional clustering effectiveness of the MEMC algorithm.Specifically, on the Reuter dataset, the MEMC algorithm achieved an F index value that was 34.9% higher than that of K-means and 17% higher than that of AP. Similarly, on the Newsgroups dataset, the MEMC algorithm exhibited an average F index value that was 27.5% higher than that of K-means and 8.8% higher than that of AP.It is important to note that while the clustering results on the Newsgroups dataset slightly lagged behind those on the Reuter dataset, this can be attributed to subtopics within certain text categories.

These subtopics belong to larger categories and contain highly similar document content. Consequently, accurately clustering such data is more challenging compared to other datasets. Furthermore, the study observed that incorporating domain knowledge to guide the clustering process further enhanced its effectiveness, especially when analyzing specific datasets. By leveraging domain knowledge, the MEMC algorithm could leverage prior information to improve clustering results.These findings emphasize the superiority of the MEMC algorithm over traditional clustering algorithms like K-means and AP. The MEMC algorithm's ability to handle diverse datasets, reach peak performance faster with increasing category features, and benefit from domain knowledge make it a valuable tool for various clustering tasks in real-world applications.

This study stands out due to its specific findings on the Reuter dataset, where the HIDDENjnt algorithm demonstrated an impressive advantage of 22% compared to other algorithms. Notably, this improvement was observed consistently across the dataset, with a minimum improvement of 10%. On the other hand, the HiLAP algorithm yielded optimal results with a 14% improvement. However, it should be noted that the initial attachment performance of HiLAP was relatively poor, showing a decrease of −8%. Compared to the MAGNET algorithm, the proposed MEMC algorithm did not outperform it in terms of results. The two algorithms had a similar optimal outcome, as indicated by the comparative analysis. For a more detailed understanding of these observations, please refer to Figure 1(a). These nuanced findings highlight the strengths and weaknesses of each algorithm, showcasing the varying degrees of improvement and efficiency across different datasets and evaluation metrics. By presenting a comprehensive analysis of multiple algorithms, this study provides valuable insights into the performance of each approach, facilitating a better understanding of their capabilities and limitations in clustering tasks.

What sets this observation apart is the noteworthy performance of the HIDDENjnt algorithm specifically on the Newsgroups dataset. Results show that the HIDDENjnt algorithm exhibits a significant advantage of 12% compared to other algorithms. It is worth mentioning that even the minimum improvement achieved by HIDDENjnt is −4%, indicating that it consistently outperforms other algorithms across the dataset.Similarly, the HiLAP algorithm demonstrates an optimal effect of 15% on the Newsgroups dataset, accompanied by a relative performance increase of 4%. This improvement highlights the effectiveness of the HiLAP algorithm in achieving accurate clustering outcomes on this dataset. However, when comparing the results obtained from the MAGNET algorithm on the Reuter dataset, it is noteworthy that the MAGNET algorithm in this study yields relatively superior results on the Newsgroups dataset. The most optimal outcome achieved by the MAGNET algorithm reaches 19%, indicating its strong performance in this context.For more detailed information, please refer to Figure 1(b) for more detailed information on these findings. This analysis provides valuable insights into the performance differences among algorithms, emphasizing the strengths and weaknesses of each approach when applied to specific datasets. These observations contribute to a better understanding of the effectiveness of various algorithms in clustering tasks and shed light on their potential application areas.



(a) F^{μ} on Reuter data set.

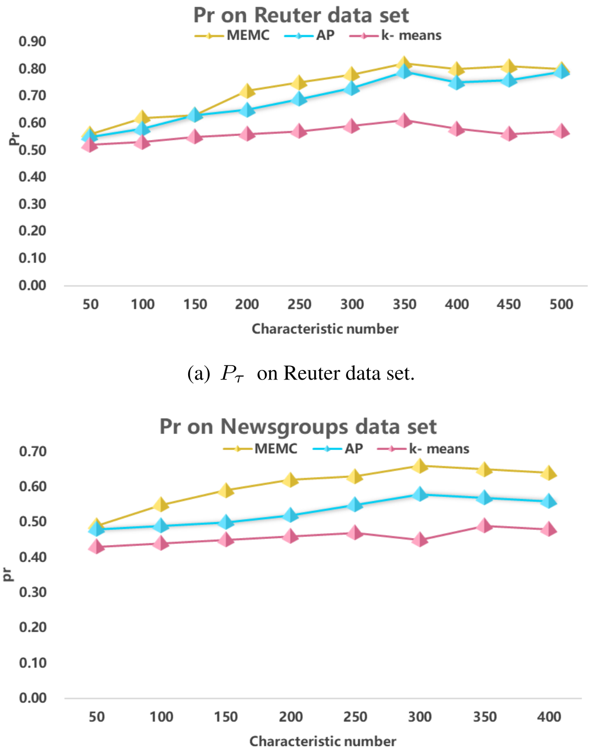
(b) F^{μ} on Newsgroups data set.

FIGURE 1. Comparison of latitude index of clustering algorithm.

To gain further insights into the clustering effect of the MEMC algorithm,we conducted additional analysis and depicted the findings in Figure2. This figure explores the variations in the overall clustering purity index of the three algorithms across different numbers of strong category features.

Based on Figure 2, it is observed that the trend of variation in the purity index of the three algorithms on the two datasets under different numbers of strong category features is consistent with the curve variation form of the micro-average F_1 index in Figure 1. In Figure2(a), the MEMC algorithm's clustering purity index value on the Reuter corpus is on average 21.7% higher than that of K-means and 4.9% higher than that of AP. In Figure 2(b), the MEMC algorithm's clustering purity index value on the Newsgroups corpus is on average 23.5% higher than that of K-means and 11.7% higher than that of AP. Overall,the MEMC algorithm achieves better clustering results than the K-means and AP algorithms across all feature numbers. However,compared to the micro-average F_1 index,the change trend of the purity index is slightly smoother. This is because the purity index is a weighted average that reflects the average result of the overall cluster classification. Thus,the purity index changes more smoothly than the micro-average F_1 index.

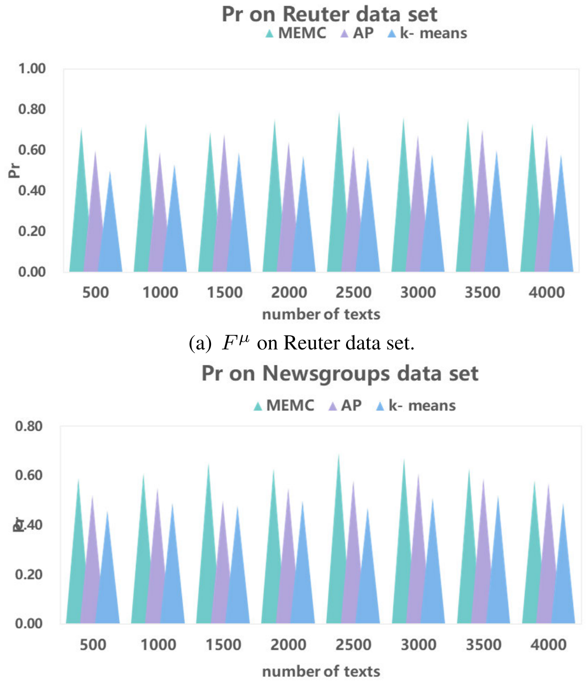
Based on the above research,it can be concluded that the MEMC algorithm can better measure text by incorporating the maximum entropy principle and a small amount of supervised prior information,thus obtaining better clustering results on different datasets. Additionally,applying domain knowledge to the clustering algorithm can improve the



(a) P_τ on Reuter data set.

(b) P_τ on Newsgroups data set.

FIGURE 2. Comparison of latitude index of clustering algorithm.



(a) F^μ on Reuter data set.

(b) P_τ on Newsgroups data set.

FIGURE 3. Comparison of latitude index of clustering algorithm.

clustering effectiveness, which can be explored further in future research.

To further test the effectiveness of each clustering algorithm,the researchers incorporated the maximum entropy principle into the MEMC,classical AP,and K-means clustering algorithms. They applied them to two text datasets of different sizes to observe the change trend of the purity P_τ value of the overall clustering results. The results indicate that,on large-scale datasets,MEMC algorithm can better measure the text aggregation effect and obtain better clustering results by introducing the maximum entropy principle,further highlighting the effectiveness and superiority of the MEMC algorithm. Furthermore,the researchers found that clustering algorithm selection and parameter settings also affect the clustering effectiveness. For instance,on the 20 Newsgroups dataset,the purity value of the MEMC algorithm increases with the number of category features. However,for other datasets,the clustering effectiveness of the MEMC algorithm may not be significantly improved with increasing category features. Therefore,when selecting and setting the parameters of the clustering algorithm,it is necessary to optimize and adjust according to the specific characteristics of the data to obtain better clustering results. Figure 3 presents the comparison results of the three clustering algorithms on two training sets of different sizes.

The results presented in Figure 3 indicate that the MEMC algorithm,which is based on the maximum entropy principle,performs admirably on two datasets with distinct characteristics: Reuter and Newsgroups. It achieves optimal clustering purity,demonstrating its effectiveness and practicality in conducting cluster analysis on text data. Compared to the K-means and classical AP algorithms,the MEMC algorithm exhibits higher clustering accuracy when applied to datasets of the same size. This superiority can be attributed to the MEMC algorithm's utilization of the maximum entropy principle,which provides a more comprehensive and precise description of data distribution during the clustering process.

In Figure 3(a), the clustering purity P_τ index value of the MEMC algorithm surpasses the K-means algorithm by 23.6% and AP by 12.3%. Similarly,in Figure 3(b), the MEMC algorithm exhibits an average clustering purity P_τ index value that is 22.3% higher than that of K-means and 11.1% higher than that of AP. These results highlight the outstanding clustering performance of the MEMC algorithm on various types of datasets. It is worth noting that the Newsgroups dataset contains particularly similar categories,making misclassification more prone compared to other datasets. Consequently,the cluster purity results obtained by the comparison algorithms on the Newsgroups dataset are inferior to those on the Reuter dataset. This finding aligns with the comparison results of the two datasets on the F index. Therefore,when dealing with text datasets featuring similar categories, selecting the appropriate algorithm for cluster analysis is imperative.

Furthermore,as the size of the training set continues to increase,the clustering purity of the AP and K-means algorithms either ceases to improve or slightly decreases.

In contrast,the MEMC algorithm consistently maintains a relatively stable clustering purity. This indicates that as the dataset size grows,the high-dimensional sparse characteristics of text data substantially amplify the complexity of clustering,resulting in a plateau or regression in the learning capability of the algorithm.

V. CONCLUSION

The study proposes a transformation incorporating the maximal entropy principle into the mean clustering algorithm, addressing the complex issue of reconstructing maximum entropy in clustering analysis. This approach provides a valuable reference scheme for practical applications in various domains, such as web text mining, information retrieval, and text clustering.The text clustering method based on neighborhood message propagation, utilizing strong category features and incorporating maximal entropy principles into the mean clustering algorithm, exhibits wide applicability in high-dimensional sparse non-Euclidean space problems. However, it is crucial to carefully select clustering algorithms and tune their parameters based on specific datasets and tasks to achieve accurate and stable clustering results.

Future research should focus on enhancing the scalability and efficiency of the MEMC algorithm by exploring advanced techniques, enabling its application to larger and more complex datasets. Additionally, comparing its performance with state-of-the-art text clustering algorithms will establish its superiority and identify areas for improvement.The computational complexity of the MEMC algorithm and its robustness to noise or outliers in the data need further investigation. Understanding the computational requirements and ensuring robustness are crucial for practical applications, especially when dealing with large-scale datasets or real-time systems.

The study presents the MEMC algorithm as an effective solution for text clustering, showcasing the benefits of incorporating maximal entropy principles. By addressing the research opportunities mentioned above, text clustering can advance its understanding and application, empowering information organization and retrieval in diverse domains. The proposed algorithm and findings significantly contribute to developing text clustering techniques and lay the foundation for future advancements in this field.

REFERENCES

[1] A. K. Jain, M N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[2] J. Wu, S. Pan, J. Jiang, Z. Cai, B. Du, Y. Tian, S. Wang, and H. Wang, "IEEE access special section editorial: Advanced data analytics for large-scale complex data environments," *IEEE Access*, vol. 7, pp. 33778–33786, 2019.

[3] X. Qian, A. Yuemaier, W. Yang, X. Chen, L. Liang, S. Li, W. Dai, and Z. Song, "A self-organizing multi-layer agent computing system for behavioral clustering recognition," *Sensors*, vol. 23, no. 12, p. 5435, Jun. 2023.

[4] Z. Zhang and M. Jin, "AOMC: An adaptive point cloud clustering approach for feature extraction," *Scientific Program.*, vol. 2022, pp. 1–13, Aug. 2022.

[5] M. Dai, X. Feng, H. Yu, and W. Guo, "A migratory behavior and emotional preference clustering algorithm based on learning vector quantization and Gaussian mixture model," *Appl. Intell.*, vol. 52, no. 15, pp. 17185–17216, 2022.

[6] F. Zhao, J. Fan, H. Liu, R. Lan, and C. W. Chen, "Noise robust multiobjective evolutionary clustering image segmentation motivated by the intuitionistic fuzzy information," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 2, pp. 387–401, Feb. 2019.

[7] T. A. Runkler and J. C. Bezdek, "Alternating cluster estimation: A new tool for clustering and function approximation," *IEEE Trans. Fuzzy Syst.*, vol. 7, no. 4, pp. 377–393, Aug. 1999.

[8] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping Multidimensional Data: Recent Advances in Clustering*. Cham, Switzerland: Springer, 2006, pp. 25–71.

[9] K.-L. Du, "Clustering: A neural network approach," *Neural Netw.*, vol. 23, no. 1, pp. 89–107, Jan. 2010.

[10] Y. Chen, L. Zhou, S. Pei, Z. Yu, Y. Chen, X. Liu, J. Du, and N. Xiong, "KNN-BLOCK DBSCAN: Fast clustering for large-scale data," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 6, pp. 3939–3953, Jun. 2021.

[11] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "DBSCAN: Past, present and future," in *Proc. 5th Int. Conf. Appl. Digit. Inf. Web Technol. (ICADIWT)*, Feb. 2014, pp. 232–238.

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996, vol. 96, no. 34, pp. 226–231.

[13] M. Li, X. Bi, L. Wang, and X. Han, "A method of two-stage clustering learning based on improved DBSCAN and density peak algorithm," *Comput. Commun.*, vol. 167, pp. 75–84, Feb. 2021.

[14] T. Li, S. Ma, and M. Ogihara, "Entropy-based criterion in categorical clustering," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, p. 68.

[15] M. K. Giri and S. Majumder, "On eigenvalue-based cooperative spectrum sensing using feature extraction and maximum entropy fuzzy clustering," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 8, pp. 10053–10067, Aug. 2023.

[16] S. D. Jagtap and M. P. B. Ramudu, "Cryptography based on artificial neural network," *Int. J. Appl. Innov. Eng. Manag.*, vol. 4, no. 8, p. 1, Aug. 2015.

[17] K. Oliver et al., "Methods and systems for performing high volume searches in a multi-tenant store," U.S. Patent 8 666 974, Mar. 4, 2014.

[18] X. Tan, L. Ge, T. Zhang, and Z. Lu, "Preservation of DNA for data storage," *Russian Chem. Rev.*, vol. 90, no. 2, pp. 280–291, Mar. 2021.

[19] A. K. Jain, M. N. Myrthy, and P. J. Flynn, "Data clustering: A survey," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[20] M. H. Ahmed, S. Tiun, N. Omar, and N. S. Sani, "Short text clustering algorithms, application and challenges: A survey," *Appl. Sci.*, vol. 13, no. 1, p. 342, 2022.

[21] G. Hamerly and C. Elkan, "Learning the K in K-means," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17, 2004, pp. 1–8.

[22] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *Wiley Interdiscipl. Rev. Data Mining Knowl. Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[23] L. Zhao, G. Shi, and J. Yang, "An adaptive hierarchical clustering method for ship trajectory data based on DBSCAN algorithm," in *Proc. IEEE 2nd Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2017, pp. 329–336.

[24] Y. G. Jung, M. S. Kang, and J. Heo, "Clustering performance comparison using K-means and expectation maximization algorithms," *Biotechnol. Biotechnological Equip.*, vol. 28, no. sup1, pp. S44–S48, Nov. 2014.

[25] A. Karim, "A comprehensive framework of machine learning model for clustering ham and spam emails based on unsupervised learning," Ph.D. thesis, College Eng., IT, Environ. Charles Darwin Univ. Australia, Casuarina, NT, Australia, 2021.

[26] X. Li, X. Fan, and X. Lu, "Modified fuzzy clustering algorithm based on non-negative matrix factorization locally constrained," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 8, pp. 11373–11383, Aug. 2023.

[27] J. K. C. Revanna, N. Niture, and N. Y. B. Al-Nakash, "Analysis of data using hybridized K-means clustering with PSO-JAYA algorithm," in *Proc. Int. Conf. Adv. Electron., Commun., Comput. Intell. Inf. Syst. (ICAECIS)*, Apr. 2023, pp. 423–428.

[28] D.-B. Sheng, S.-B. Kim, T.-H. Nguyen, D.-H. Kim, T.-S. Gao, and H.-K. Kim, "Fish injured rate measurement using color image segmentation method based on K-means clustering algorithm and Otsu's threshold algorithm," *J. Korea Soc. Power Syst. Eng.*, vol. 20, no. 4, pp. 32–37, Aug. 2016.

[29] H. V. Shashidhar and S. Varadarajan, "Customer segmentation of bank based on data mining–security value based heuristic approach as a replacement to K-means segmentation," *Int. J. Comput. Appl.*, vol. 19, no. 8, pp. 13–18, 2011.

[30] H. Huang, R. Zhang, F. Xiong, F. Makedon, L. Shen, B. Hettleman, and J. Pearlman, "K-means+ method for improving gene selection for classification of microarray data," in *Proc. IEEE Comput. Syst. Bioinf. Conf. Workshops (CSBW)*, 2005, p. 2.

[31] Z. Chen and Y. F. Li, "Anomaly detection based on enhanced DBScan algorithm," *Proc. Eng.*, vol. 15, pp. 178–182, 2011.

[32] C. Guan, K. K. F. Yuen, and Q. Chen, "Towards a hybrid approach of K-means and density-based spatial clustering of applications with noise for image segmentation," in *Proc. IEEE Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Social Comput. (CPSCom) IEEE Smart Data (SmartData)*, Jun. 2017, pp. 396–399.

[33] R. Kaur and S. Singh, "A survey of data mining and social network analysis based anomaly detection techniques," *Egyptian Informat. J.*, vol. 17, no. 2, pp. 199–216, Jul. 2016.

[34] S. Chatterjee, A. Maheshwari, G. Ramakrishnan, and S. N. Jagaralpudi, "Joint learning of hyperbolic label embeddings for hierarchical multi-label classification," 2021, *arXiv:2101.04997*.

[35] Y. Mao, J. Tian, J. Han, and X. Ren, "Hierarchical text classification with reinforced label assignment," 2019, *arXiv:1908.10419*.

[36] A. Pal, M. Selvakumar, and M. Sankarasubbu, "Multi-label text classification using attention-based graph neural network," 2020, *arXiv:2003.11644*.

[37] I. Csiszar, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, Feb. 1975.

[38] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, no. 9, pp. 939–952, Jun. 1982.

[39] P. Marsh, "A two-sample nonparametric likelihood ratio test," *J. Nonparametric Statist.*, vol. 22, no. 8, pp. 1053–1065, Nov. 2010.

[40] N. Tatti, "Computational complexity of queries based on itemsets," 2019, *arXiv:1902.00633*.

[41] J. Xu, W. Zhang, and R. Sun, "Efficient reliability assessment of structural dynamic systems with unequal weighted quasi-Monte Carlo simulation," *Comput. Struct.*, vol. 175, pp. 37–51, Oct. 2016.

[42] M. Singh and N. K. Vishnoi, "Entropy, optimization and counting," 2013, *arXiv:1304.8108*.

[43] S. Dalleiger and J. Vreeken, "The relaxed maximum entropy distribution and its application to pattern discovery," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2020, pp. 978–983.

[44] S. Li and R. Cai, "The generalized maximum belief entropy model," *Soft Comput.*, vol. 26, no. 9, pp. 4187–4198, May 2022.

[45] C. Carretero-Campos, P. Bernaola-Galván, A. V. Coronado, and P. Carpena, "Improving statistical keyword detection in short texts: Entropic and clustering approaches," *Phys. A, Stat. Mech. Appl.*, vol. 392, no. 6, pp. 1481–1492, Mar. 2013.

[46] A. Singhal and D. K. Sharma, "Keyword extraction using Renyi entropy: A statistical and domain independent method," in *Proc. 7th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, vol. 1, Mar. 2021, pp. 1970–1975.

XUMIN ZHAO is currently pursuing the Ph.D. degree with Philippine Christian University. She serves as an Adjunct Researcher with the Key Laboratory of Open Data of Zhejiang Province and is a faculty member at Zhejiang Yuexiu University. Her primary research interests lie in the exploration of artificial intelligence algorithms and their applications in business.

GUOJIE XIE is currently serves as the Head of External Affairs at the Key Laboratory of Open Data, Zhejiang Province. He has extensive experience in digital transformation, digital government, and industrial internet field research, and has gained significant practical experience in project delivery, solution customization, and intellectual property protection. In addition, his vast experience in managing major projects for government and central state-owned enterprises has equipped him with the skills necessary to lead the development of innovative field products.

YI LUO has published numerous articles in multiple SCI and SSCI journals. His main research interests include artificial intelligence, complex networks, and computational systems.

FENGHUA LIU received the master's degree from Xi'an Jiaotong University. She is currently an Associate Professor, mainly in big data and artificial intelligence research. She is a Referee with the National Vocational College Skills Competition and an Expert with the Teacher Teaching Ability Competition. She is also a Municipal Science and Technology Expert Database Member and an Expert with the Municipal 1+X Certificate Learning Achievement Professional Certification Committee.

HONGPENG BAI received the master's degree from Changchun University of Science and Technology, China, in 2021. He is currently pursuing the Ph.D. degree with the College of Intelligence and Computing, Tianjin University, China. His current research interests include malware detection and the IoT security.