Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8881112, 16 pages
https://doi.org/10.1155/2020/8881112

Research Article

High-Dimensional Text Clustering by Dimensionality Reduction and Improved Density Peak

Yujia Sun [1,2] **and Jan Platoš** [1]

[1]*Department of Computer Science, Technical University of Ostrava, 17.listopadu 2172/15, Poruba, Ostrava 70800, Czech Republic*
[2]*Institute of Network Information Security, Hebei GEO University, No. 136 East Huai'an Road, Shijiazhuang Hebei 050031, China*

Correspondence should be addressed to Yujia Sun; yujia.sun.st@vsb.cz

Received 30 May 2020; Revised 21 September 2020; Accepted 20 October 2020; Published 28 October 2020

Academic Editor: Chao-Yang Lee

This study focuses on high-dimensional text data clustering, given the inability of K-means to process high-dimensional data and the need to specify the number of clusters and randomly select the initial centers. We propose a Stacked-Random Projection dimensionality reduction framework and an enhanced K-means algorithm DPC-K-means based on the improved density peaks algorithm. The improved density peaks algorithm determines the number of clusters and the initial clustering centers of K-means. Our proposed algorithm is validated using seven text datasets. Experimental results show that this algorithm is suitable for clustering of text data by correcting the defects of K-means.

1. Introduction

Clustering is the main technique used for unsupervised information extraction. In clustering, the aim is to divide the unlabelled dataset into multiple nonoverlapping class clusters, making the data points in the cluster as similar as possible, while making the data points between the clusters as different as possible. In text clustering, text vectors are characterized by high dimension, sparsity, and correlation among dimensions, which requires improvements to the clustering algorithm to process high-dimension text [1, 2].

When the K-means method is used to process high-dimensional data, the "Curse of Dimensionality" [3] problem becomes prominent, and the redundancy index also increases. Consequently, the conventional clustering method cannot process the data accurately. Some research [4–9] has proposed improvements on the text clustering algorithm, and some studies [10, 11] have proposed improvements on the K-means algorithm. To apply the K-means, it is necessary to specify the number of clusters in advance and randomly select the initial clustering centers. The clustering result is greatly influenced by the selection of the initial center point. Improper selection of the initial center can easily cause the clustering result trap into the local optimal solution and lead to an inaccurate clustering result.

In recognition of these problems, we propose an enhanced K-means text clustering algorithm based on the clustering by fast search and find of density peaks (DPC) algorithm [12]. Since text-based data is usually high-dimensional and sparse, we propose a deep random projection dimensionality reduction framework, named Stacked-Random Projection (SRP), a greedy layer-wise architecture. We first use the dimensionality reduction method to reduce the dimension of the high-dimensional text feature vectors. Then use the improved density peaks algorithm to determine the number of clusters and the initial clustering centers, after which the K-means algorithm is used for clustering.

The organization of this paper is as follows. The proposed methodology is discussed in Methods. In Experiments and Discussion, experimental results are explained. Finally, Conclusions concludes the paper and highlights future work related to the study.

2. Methods

2.1. Stacked-Random Projection. The basic idea of random projection is to choose a random hyperplane to map original

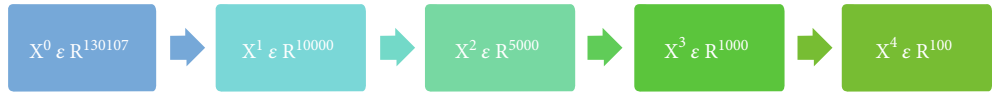

Figure 1: The SRP dimensionality reduction process for the 20-newsgroups dataset. The 4-layer SRP realizes the dimensionality reduction, using the random projection method from 130,107 down to 10k, down to 5k, down to 1k, and down to 100.

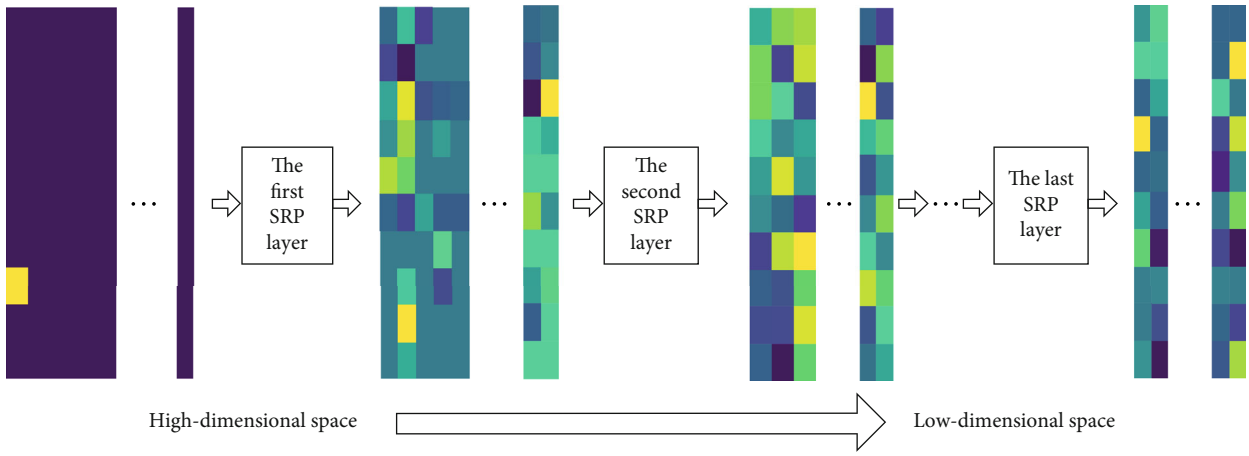

Figure 2: Architecture of Stacked-Random Projection.

variables into a low-dimensional space. In 1948, Johnson and Lindenstrauss proposed a theorem, nowadays termed the Johnson-Lindenstrauss lemma (JL) [13]. JL lemma is the theoretical basis of random projection, which guarantees that the subspace errors generated by random projection are controllable. The JL lemma states that for any $0 < \varepsilon < 1$, and any integer n, let k be a positive integer such that

$$k \geq \left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}\right)^{-1} \ln n. \tag{1}$$

Then, for any n-point set $\mathbf{V}$ in $\mathbf{R}^d$, there is a map f: $\mathbf{R}^d \longrightarrow \mathbf{R}^k$, such that for all $u, v \epsilon \mathbf{V}$,

$$(1\text{-}\varepsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \varepsilon)\|u - v\|^2. \tag{2}$$

It indicates that by using random projection, the original high-dimensional data is reduced to low-dimensional data, and the distance between the original data is maintained approximately with a high probability. Zhang et al. [14] proposed a random projection ensemble approach and applied it to the prediction of drug-target interaction. Gondara [15] also proposed an ensemble random projection, in which the random projection matrix is applied to different subsets of the original dataset, and which can achieve greater classification accuracy compared with the random forest and AdaBoost methods.

According to the Johnson-Lindenstrauss lemma, the minimum size of the target dimension after dimensionality reduction that guarantee the ε embedding is given by Equation (3):

$$\text{dimension} \geq \frac{4 \log\left(n_{\text{samples}}\right)}{\left(\varepsilon^2/2 - \varepsilon^3/3\right)}. \tag{3}$$

For example, where n_{samples} is the number of samples, it would require at least 6,515 dimensions to project 2k samples without too much distortion ($\varepsilon = 0.1$). Thousands of dimensions are still high-dimensional data for the following step such as classification or clustering. Inspired by stacked Auto-Encoder, we propose a deep random projection framework, named Stacked-Random Projection (SRP), which incorporates random projection as its core stacking element. The SRP framework with k layers uses the input data as the first layer, and the output of the lth ($l < k$) layer is taken as the $(l + 1)$ layer input. In this way, a group of random projections method can be combined layer by layer in a stack.

The main idea of the SRP dimensionality reduction method based on the high-dimensional text feature vector can be illustrated by means of taking the 20-newsgroups dataset as an example (further details are provided in Experiments and Discussion). First, the dataset is subjected to tokenization, stop-words removal, and TF-IDF in order to obtain the high-dimensional sparse text vector space (the feature dimension of the 20-newsgroups dataset was found to be 130,107). Then, a 4-layer SRP is constructed, this process is shown in Figure 1. Thus, the dimensionality reduction process from high dimensionality to low dimensionality is completed. The illustration of our proposed SRP is provided in Figure 2.

2.2. Improved DPC. The DPC algorithm is a granular computing model based on two assumptions: (1) the clustering center is surrounded by neighbour data points with lower local density; (2) the distance between any clustering center and data points with higher density is relatively far. In recent years, DPC has been applied in many fields, particularly natural language processing, due to its process and its effectiveness. The DPC algorithm can cluster data of different

dimensions and shapes. At present, many researchers have researched DPC and have also proposed many improved algorithms. The main optimization aspects are speed improvement [16], accuracy improvement [17–19], and other aspects [20, 21]. Heimerl et al. [22] applied the DPC algorithm in the high-dimensional space to estimate the optimal cluster numbers for a given set of documents and assigned stability to one of the peaks based on the density structure of the data; however, the resulting computing speed of the DPC algorithm in the high-dimensional space was slow. Wang et al. [23] used DPC to measure the hierarchical relevance and diversity of sentences and selected highly representative sentences to generate news summaries. However, they reported that if there are multiple peaks in the sentence, then the key sentence will be redundant.

For any point i, two properties of the local density and relative distance are required. The calculation of these two attributes depends on the distance between any two points v_i and v_j in the graph. The two attributes are defined as follows:

Definition 1. local density ρ_i (Gaussian kernel):

$$\rho_i = \sum_j e^{-\left(\frac{d_{ij}}{d_c}\right)^2}, \tag{4}$$

where d_{ij} is the Euclidean distance between v_i and v_j, and d_c is the cut-off distance; these are important parameters for calculating ρ_i. One recommended practice is to select d_c so that the average nearest neighbour from each point is 1%~2% of the total dataset size. As can be seen in Equation (4), the more points i contained in d_c, the greater the local density ρ.

Of the text clustering methods, the K-means method based on cosine similarity is still the most widely used text clustering algorithm due to its simplicity and fast convergence [24]. For text vectors, using cosine similarity has a better effect than Euclidean distance. Euclidean distance is a direct measure of the linear interval or length between vectors and is an absolute value of the difference in dimensional values. Cosine similarity describes the similarity between vectors using the cosine value of the angle, that is, the direction, and pays more attention to the difference between the relative levels of the dimensions. In text similarity analysis, one feature of similarity is the occurrence of the same words at the same time, which translates into nonzero values for the same dimension at the same time. We therefore redefine Definition 1 in terms of cosine similarity.

Definition 2. Local density ρ_i based on cosine similarity (Gaussian kernel):

For any two vectors in space $v_i = (x_1, x_2, \cdots, x_n)$ and $v_j = (y_1, y_2, \cdots, y_n)$, the cosine similarity is defined as the cosine of the angle between the two vectors:

$$\cos(i,j) = \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k^2}\sqrt{\sum_{k=1}^{n} y_k^2}} = \frac{\sum_{k=1}^{n} x_k y_k}{\|x\| \bullet \|y\|}, \tag{5}$$

$$\rho_i = \sum_j e^{-\left(\frac{\cos(i,j)}{\cos_c}\right)^2}, \tag{6}$$

where $\cos(i,j)$ is the cosine similarity between v_i and v_j, and $\cos_c$ is the cut-off distance which needs to manually set the value to the nearest neighbour number of the sample approximately 1%~2% of the size of the entire dataset. As can be seen in Equation (6), the more points i contained in $\cos_c$, the greater the local density ρ.

Definition 3. Relative distance δ_i:

$$\delta_i = \begin{cases} \max\limits_j \cos(i,j) & \rho_i \text{is the maximum} \\ \min\limits_{j:\rho_i<\rho_j} \cos(i,j) & \text{otherwise.} \end{cases} \tag{7}$$

Equation (7) indicates that cosine similarity distance δ_i can be obtained by calculating the minimum distance from the data point x_i to any point with a density greater than that. After calculating the two parameters, a decision graph with ρ as the horizontal axis and δ as the vertical axis can be constructed. By observing the decision graph, the decision graph divides the data points into three different types, namely the density peak point, the normal point, and the outlier point. As shown in Figure 3, the data points are arranged in the order of decreasing density. There are five points that stand out, which are spread out towards the upper right corner of the decision graph, with varying high ρ values and higher δ values. These five points indicate that there are no data points with higher density than these five points in a larger area. Therefore, these five points are the so-called peak density points, and so they make a suitable clustering center. In order to better verify the accuracy of the clustering center point in the decision graph, the DPC define another variable $\gamma = \rho * \delta$, where a clustering center point has a large ρ value and δ value, the clustering center has a higher γ value. We conclude from our analysis of the decision graph and that ρ and δ are of two different orders of magnitude. To avoid the influence of different orders of magnitude, it is necessary to normalize them.

$$\rho_i' = \frac{\rho_i - \rho_{\min}}{\rho_{\max} - \rho_{\min}}, \tag{8}$$

$$\delta_i' = \frac{\delta_i - \delta_{\min}}{\delta_{\max} - \delta_{\min}}, \tag{9}$$

$$\gamma_i = \rho_i' \delta_i'. \tag{10}$$

The γ values for Equation (10) are plotted in Figures 4 and 5. Figure 4 verifies the correctness of the clustering centers in the decision graph shown in Figure 3. Figure 5 is plotted according to the descending order of γ value, and it can be noted that γ value changes from large to small. The point at the clustering center has an enormous γ value, while the noncenter point has a smaller γ value, and the change tends to be flat. It can be concluded that according to the γ value, there are five clustering centers.

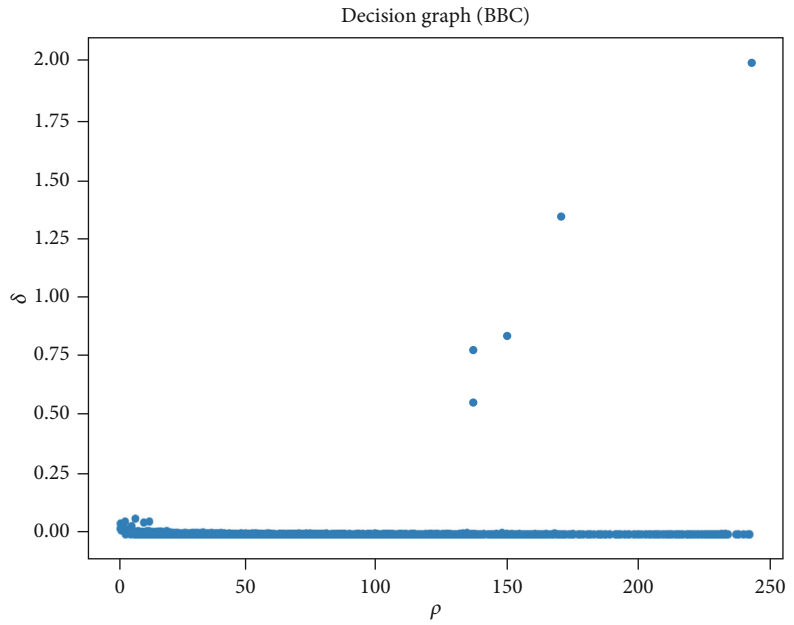

Figure 3: The decision graph of BBC dataset.

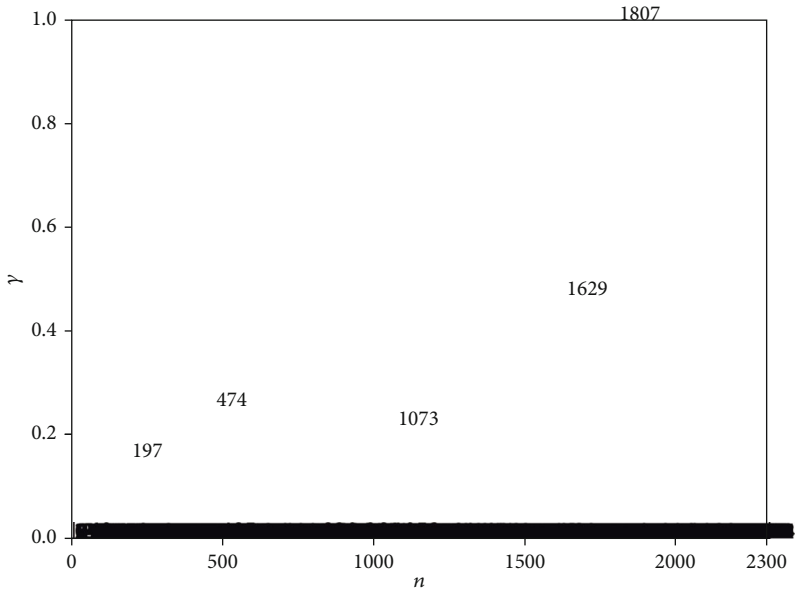

Figure 4: The value of γ according to the $\rho * \delta$ in Figure 3.

2.3. DPC-K-means. The K-means clustering algorithm cannot extract data features effectively when processing high-dimensional data directly, and problems also occur when it randomly selects initial clustering centers and specifies the number of clustering in advance. These problems have been researched in numerous papers over the recent decades, as discussed elsewhere [25–27]. Therefore, we propose an improved method using the DPC algorithm.

We first use SRP or random projection to reduce the dimensionality of high-dimensional text data and then combine it with the improved DPC algorithm. The choice of dimensionality reduction method SRP or random projection depends on whether the feature vector dimension is greater than the target dimension calculated according to Formula (3). If the feature vector dimension is greater than the minimum size of the target dimension, the SRP dimension reduction framework is performed. If the feature vector dimension is less than or equal to the target dimension, random projection is used directly. Using the cosine similarity calculation of ρ and δ, we select some points with high local density, which are far apart from each other as the clustering center; by doing so, the initial clustering center and the number of clusters can be obtained automatically; this makes the clustering algorithm, which we name the DPC-K-means. The improved algorithm is described below:

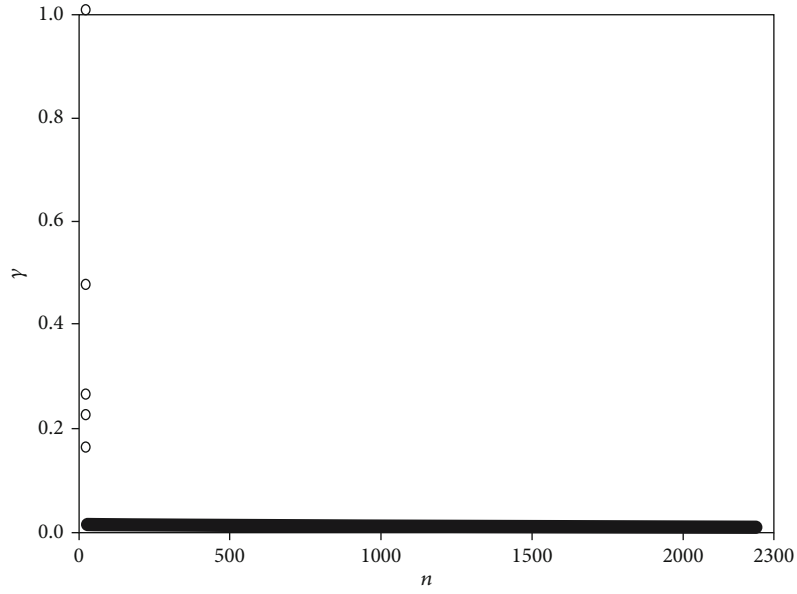

Figure 5: The value of γ of Figure 4 in decreasing order.

The DPC-K-means.
Input: text feature vector $\mathbf{A} \epsilon \mathbf{R}^{n\times d}$, t is the minimum size of the target dimension.
Output: the clustering results.
Begin:
Step1: determine whether d is greater than t calculated according to Formula (3). If d is greater than t, use the SRP dimension reduction framework in Step2. If d is less than or equal to t, random projection is used in Step2.
Step2: the SRP dimension reduction framework is used to reduce the dimensionality of $\mathbf{A}$ layer by layer, until matrix $\mathbf{A}^{'}$ after dimension reduction is obtained. Or directly use random projection to reduce the dimension to get the matrix $\mathbf{A}^{'}$.
Step3: Calculate the ρ value and δ value of $\mathbf{A}^{'}$ according to Equations (6) and (7) and plot the decision graph with ρ and δ axes.
Step4: calculate the γ value according to Equation (10) to verify the clustering centers and the number of clusters.
Step5: perform K-means clustering: the clustering centers obtained in Step4 are used as the initial cluster centers, and the number of clusters is used as the k value for K-means clustering.

Algorithm 1:

Suppose n input data, the original dimension $d, t^{'}$ is the dimension of implementing SRP or random projection to reduce dimension to low-dimensional space, and the time complexity analysis of DPC-K-means algorithm is as follows:

(1) The time complexity of a single random projection in Step2 is $O(ndt^{'})$. The time complexity of Stacked-Random Projection is $O(ndh + nhl + \cdots + nrt^{'})$ (l is the target dimension of the second layer, and r is the target dimension for the penultimate layer)

(2) The time complexity of Step3 is to calculate ρ and δ, which is $O(n^2)$

(3) The time complexity of Step4 is to calculate γ and sort γ in descending order, which is $O(n \log_2 n)$

(4) The time complexity of Step5 K-means for specifying the cluster center and the number of clusters is $O(knt^{'})$.

The total time complexity of DPC-K-means algorithm is $O(n^2)$.

Figure 6 shows the overall structure of the proposed method. Table 1 shows the time complexity of several clustering algorithms.

3. Experiments and Discussion

3.1. Summarization Datasets. Experimental work was conducted on seven standard text datasets. The summary of datasets is presented in Table 2. Datasets are described as follows. The features are obtained by tokenization, stop-words removal, and TF-IDF.

The BBC news dataset (http://mlg.ucd.ie/datasets/bbc.html.) has a total of 2,225 text files on five topical areas published on the BBC news website. Text documents were arranged into folders containing five labels: business, entertainment, politics, sports, and technology.

The 20-newsgroups dataset (http://scikit-learn.org/stable/modules/classes.html#module-sklearn.datasets) of approximately 20k newsgroup documents was partitioned evenly across the 20 different newsgroups. We selected 1k

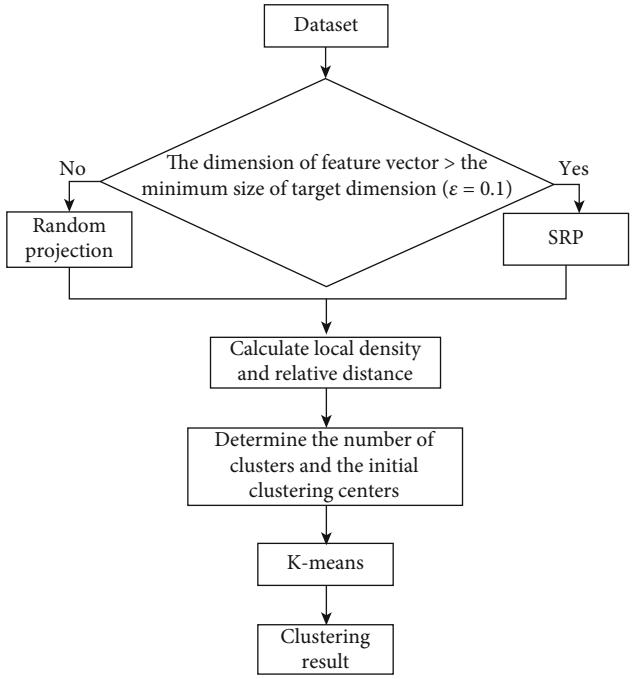

Figure 6: The detailed process of DPC-K-means.

documents and 4~8 various newsgroups (4 groups~8 groups) for our experimental dataset.

The Sports Article dataset (http://archive.ics.uci.edu/ml/datasets.php) was labelled using Amazon Mechanical Turk as objective or subjective.

The Asian Religious (http://archive.ics.uci.edu/ml/datasets.php) dataset was the words from the bag of words preprocessing of the mini-corpus made up of eight religious books.

The CNAE-9 dataset (http://archive.ics.uci.edu/ml/datasets.php) contains 1,080 documents of free text business descriptions of Brazilian companies which were categorized into a subset of nine categories.

The Stack Overflow dataset (http://www.kaggle.com/c/predict-closed-questions-on-stack-overflow/download/train.zip) is challenge data published on http://Kaggle.com/. The dataset consists of 3,370,538 samples dated from July 31, 2012, to August 14, 2012. In our experiments, we randomly selected 167 question titles from 4 different tags.

The Amazon dataset (http://archive.ics.uci.edu/ml/datasets.php) is the product reviews extracted from websites and marked with positive and negative.

3.2. Simulation Environments. The simulation environments for all algorithms performed in our experiments were as follows: the Python 3.7 software environment running with Intel i7-7500U CPU, 2.70GHz with 8GB RAM.

3.3. Experiment 1. According to Formula (3), the minimum size of the target dimension ($\varepsilon = 0.1$) of the BBC and 20-newsgroups datasets is 6,609 and 5,920. According to the flowchart Figure 6, the feature vector dimensions of the two datasets are larger than the minimum size of the target dimension, so that SRP was used to reduce the dimensionality of these two datasets. We compared the dimensionality reduction performance of Principal Component Analysis (PCA), Multiple Dimensional Scaling (MDS), Random Projection (RP), and Stacked-Random Projection (SRP). To correctly compare the performance of these dimensionality reduction methods, we experimentally reduced the feature vector of the BBC news dataset and 20-newsgroups dataset to 2k, 500, and 100. Table 3 shows the run time (time), mean ratio of distances (projected/original, ratio), and the standard deviation of ratio of distances (projected/original, standard deviation). The mean ratio of distances is the degree to which the distance between the original data is maintained in the low-dimensional space when the original high-dimensional data is reduced to low-dimensional data. The value is approximately close to 1, indicating better preservation. The smaller the standard deviation of the ratio of distances, the closer is it to the mean ratio of distances. As shown in Table 3, RP and SRP considerably shorten the run time of dimension reduction compared with PCA and MDS. We can see that there is little difference in the distribution of the distortion between SRP and RP for high values of the dimension. But for low values of the dimension, the distortion distribution is controlled, and the distances are well preserved by the SRP. Text data is usually high-dimensional and small-sampling data. The characteristic of high-dimensional and small-sampling data is that the number of dimensions is much larger than the number of samples. SRP is suitable for dimensionality reduction of this type of data, which significantly reduces the running time of dimensionality reduction, and the distances are well preserved.

3.4. Experiment 2. Since DPC is a clustering algorithm, we use the Euclidean distance and cosine similarity to calculate DPC local density ρ_i and observe the difference between these two methods by clustering performance metrics. According to Formula (3), the minimum size of the target dimension ($\varepsilon = 0.1$) of the BBC and 20-newsgroups datasets is 6,609 and 5,920. According to the flowchart Figure 6, SRP was used to reduce the dimensionality of these two datasets to 100 dimensions. The minimum size of the target dimension of Sports Article, CNAE-9, and Stack Overflow datasets is larger than the feature vector dimension, so that the dimension can be reduced by random projection to 100 dimensions. The feature dimensions of the Asian Religions and Amazon datasets are ≤100, so there is no need for dimension reduction in this experiment. To correctly compare these two methods' performance, we used the four cluster evaluation metrics—ARI (Adjusted Rand Index), NMI (Normalized Mutual Information), FMI (Fowlkes-Mallows Index), and Clusters (the number of clusters)—to evaluate the performance of the clustering algorithm. ARI, NMI, and FMI are all used to measure the consistency between clustering results and real category data, among which ARI, NMI, and FMI have value ranges of [-1,1], [0,1], and [0,1], respectively. The higher the three evaluation metrics' values, the better the clustering quality, and the more consistent the clustering results are with the real category data. Clusters are the number of clusters after DPC. By comparing with Table 2, we can compare which method of the Euclidean distance and cosine similarity

Table 1: The time complexity of several clustering algorithms.

	DPC-K-means	DPC	K-means	DBSCAN	Spectral Clustering	Affinity Propagation
Time complexity	$O(n^2)$	$O(n^2)$	$O(nkt')$	$O(n^2)$	$O(n^3)$	$O(n^2 \log n)$

Table 2: The summary of datasets.

Dataset	Instances	Dimension of features	Clusters	Label
BBC	2,225	11,227	5	Yes
20-newsgroups	1000	13,0107	4~8	Yes
Sports article	1,000	348	2	Yes
Asian Religious	590	39	8	Yes
CNAE-9	1,080	857	9	No
Stack Overflow	167	167	4	Yes
Amazon	100	100	2	Yes

Table 3: The run time, ratio, and standard deviation of each dimension reduction method reduce the dimension to 2,000, 500, and 100.

		Dimension = 2,000			Dimension = 500			Dimension = 100		
		Time (s)	Ratio	Standard deviation	Time (s)	Ratio	Standard deviation	Time (s)	Ratio	Standard deviation
BBC	PCA	30.52	1.00	0.02	15.16	0.57	0.09	4.99	0.23	0.09
	MDS	57.99	1.00	0.02	28.67	1.00	0.07	15.47	1.00	0.07
	RP	1.12	1.00	0.04	0.56	1.00	0.08	0.41	1.01	0.18
	SRP	2.87	1.00	0.05	2.38	1.00	0.07	2.27	1.00	0.12
20-newsgroups	PCA	24.87	1.00	0.00	13.35	1.00	0.08	4.20	1.00	0.1
	MDS	65.27	1.00	0.02	44.74	1.00	0.03	23.13	0.99	0.06
	RP	0.10	0.99	0.06	0.46	1.00	0.12	0.31	0.99	0.28
	SRP	3.33	1.00	0.05	2.85	1.00	0.07	2.72	1.00	0.15

Table 4: The clustering performances of local density calculated by Euclidean distance and cosine similarity.

Dataset	ARI		NMI		FMI		Clusters	
	Euclidean	Cosine	Euclidean	Cosine	Euclidean	Cosine	Euclidean	Cosine
BBC	0.8422	0.9002	0.8223	0.8681	0.8759	0.9204	5	5
4 groups	0.9715	0.9781	0.9523	0.9623	0.9786	0.9836	4	4
5 groups	0.8438	0.8433	0.8411	0.8381	0.8851	0.8846	5	5
6 groups	0.6195	0.6759	0.6874	0.7351	0.7326	0.7700	6	6
7 groups	0.2213	0.5858	0.3039	0.6487	0.4999	0.7031	5	7
8 groups	0.1889	0.4664	0.2672	0.5507	0.4607	0.6138	5	8
Sports Article	0	0	0	0	0.5674	0.7298	1	2
Asian Religious	0.0562	0.0189	0.1288	0.0163	0.3145	0.4665	6	8
Stack Overflow	0	0	0	0	0.3660	0.4399	2	4
Amazon	0	0.0014	0	0.0048	0.5515	0.6696	2	2
CNAE-9	—	—	—	—	—	—	6	9

can cluster accurately. Table 4 shows the clustering performances of local density calculated by the Euclidean distance (Euclidean) and cosine similarity (Cosine).

To further judge the clustering performance proposed in this paper, a paired t-test was used to test the clustering significance. A paired t-test is used to determine whether

Table 5: Table 4's paired t-test results of ARI, NMI, and FMI.

Pairing method	Paired t-test index	ARI	NMI	FMI
Euclidean and Cosine	t	-1.70	-1.40	-4.16
	p	0.1240	0.1958	0.0025

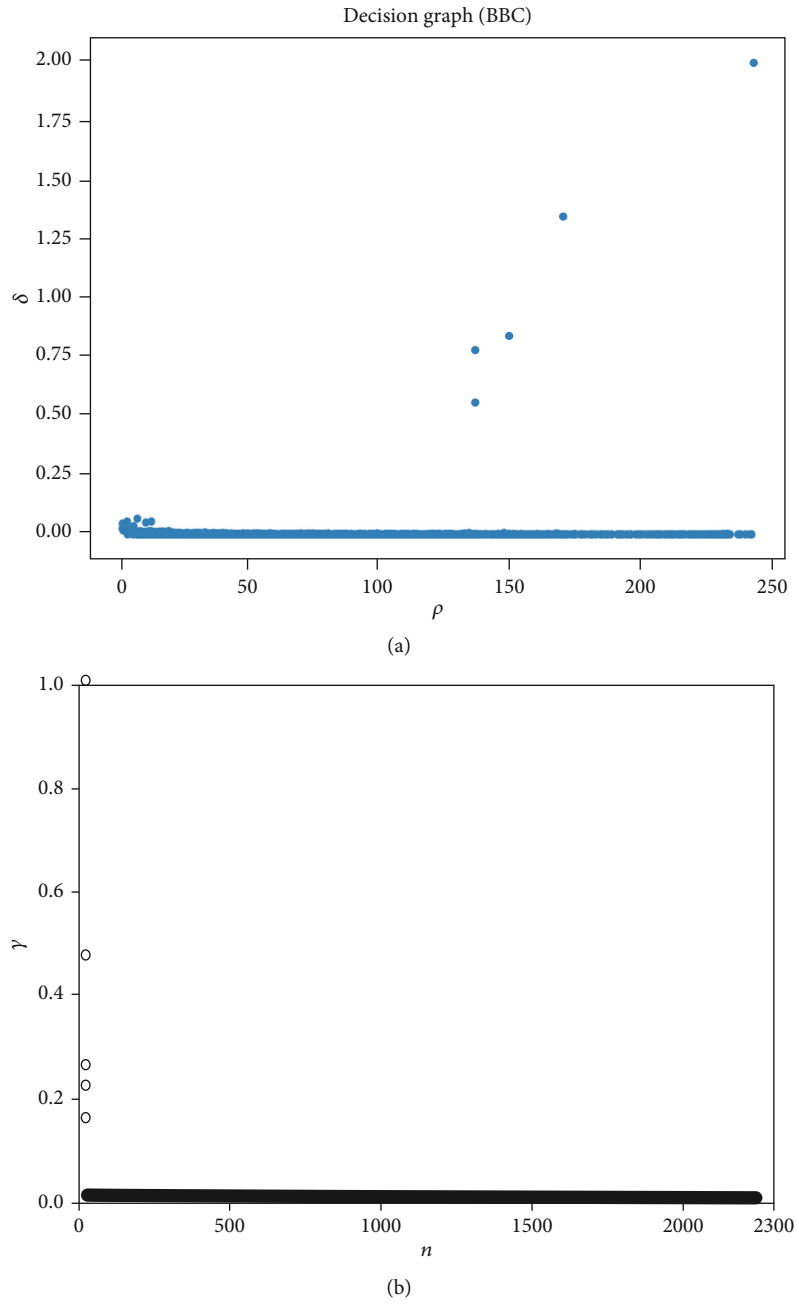

Figure 7: The improved DPC clustering of the BBC dataset. (a) Decision graph. (b) γ value.

there is a significant difference between the two samples. The Euclidean distance and cosine similarity were used to calculate the local density of DPC and test the cluster evaluation metrics. The p value gives the probability of observing the test results under the null hypothesis. The confidence level is at 95%, and the cut-off value of p is 0.05; if $p < 0.05$, the proposed algorithm clustering results and the comparison algorithm are significantly different. If $p \geq 0.05$, there is no significant difference between the proposed algorithm and the comparison algorithm's clustering results. Table 5 shows the paired t-test results of each evaluation metric of Euclidean distance (Euclidean) and cosine similarity (Cosine) in Table 4.

As shown in Table 5, there are substantial differences in FMI between the Euclidean distance and cosine similarity and no significant difference in ARI, NMI. As can be seen from the number of clusters of Tables 2 and 4, the improved DPC of the local density calculated with cosine similarity can accurately determine the number of clusters. Figure 7 shows the decision graph, the γ values of the BBC dataset following dimensionality reduction by SRP. Figure 8 shows the decision graph, the γ values of the four newsgroups in the 20-

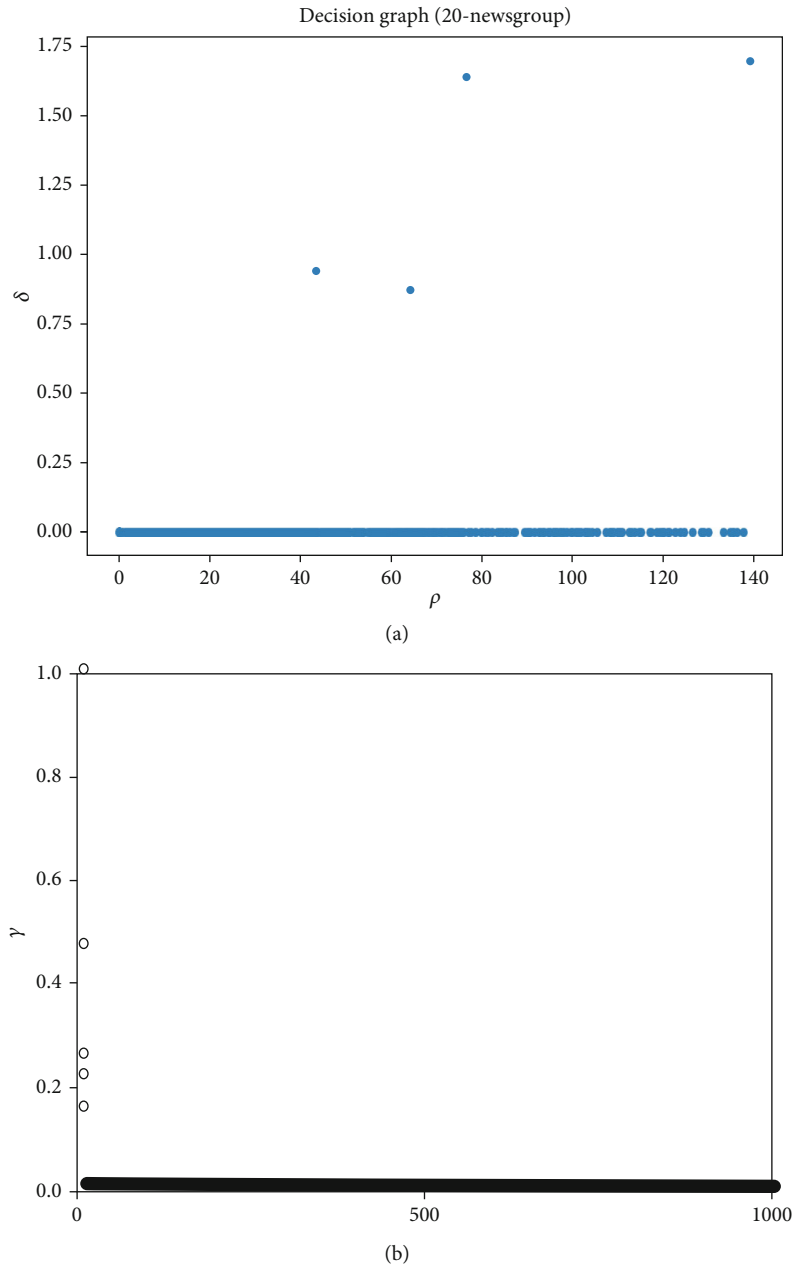

Figure 8: The improved DPC clustering of the four newsgroups in the 20-newsgroups dataset. (a) Decision graph. (b) γ value.

newsgroups dataset following dimensionality reduction by SRP. Figure 9 shows the decision graph, the γ values of the Amazon dataset. Figure 10 shows the decision graph, the γ values of the Sports Article dataset. As shown in these figures, improved DPC can accurately determine the dataset of the number of clusters, indicating that using cosine similarity to calculate the local density of DPC is better than using the Euclidean distance. Therefore, cosine similarity is more suitable for text vector calculation.

3.5. Experiment 3. We compared the clustering performance of DPC, DBSCAN, Spectral Clustering, Affinity Propagation, and DPC-K-means. In a comparative study of these clustering algorithms, we used the four evaluation metrics—ARI (Adjusted Rand Index), NMI (Normalized Mutual Information), FMI (Fowlkes-Mallows Index), and MSE (Mean Squared Error)—to evaluate the performance of the clustering algorithm. The mean-square error (MSE) is the average of the sum of squares of the difference between the predicted value and the real value used to measure the expected result. It is nonnegative, and values closer to zero are better. For even comparisons with these methods, we repeated the experiment ten times to obtain the average clustering performance as the final performance of each method. Table 6 shows the ARI of each method. Table 7 shows the NMI of each method. Table 8 shows the FMI of each method. Table 9 shows the MSE of each method.

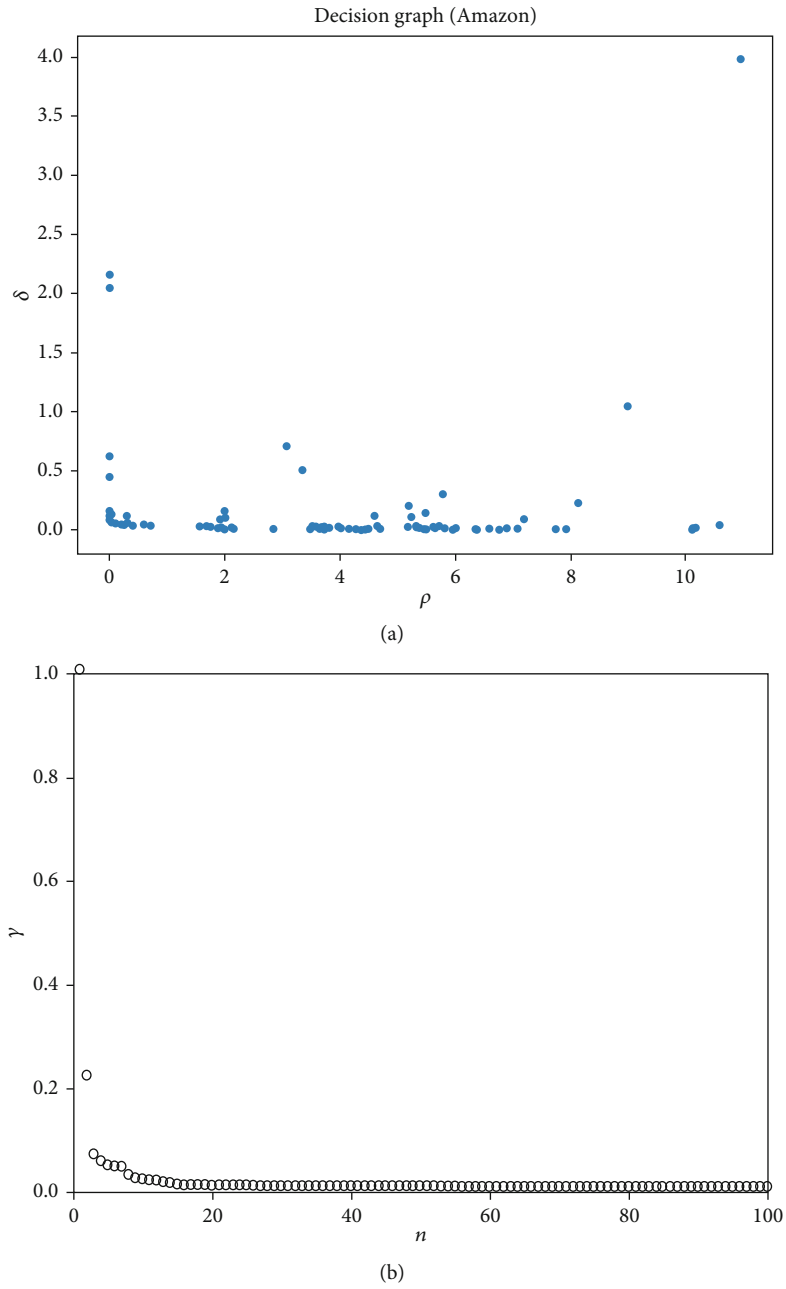

Figure 9: The improved DPC clustering of the Amazon dataset. (a) Decision graph. (b) γ value.

To further judge the difference between the clustering results of the algorithm proposed in this paper DPC-K-means and those of other cluster methods, a paired t-test was used to test the clustering results significance. Table 10 shows the paired t-test results of each evaluation metric of these methods in Tables 6–9. The p value gives the probability of observing the test results under the null hypothesis. The confidence level is at 95%, and the cut-off value of p is 0.05; if $p < 0.05$, the proposed algorithm's clustering results and the comparison algorithm are significantly different. If $p \geq 0.05$, there is no significant difference between the proposed algorithm and the comparison algorithm clustering performance.

As shown in Table 10, there are significant changes in NMI, FMI, and MSE metrics between DPC-K-means and comparison methods. DPC-K-means is superior to comparison algorithms in NMI, FMI, and MSE. DPC-K-means compared to DPC and Spectral Clustering has no significant difference in ARI, indicating that DPC and Spectral Clustering are performed as well as DPC-K-means on the ARI metric. A one-sample t-test method is used to evaluate the DPC-K-means algorithm significance on different datasets. Taking the FMI evaluation metric in the BBC dataset as an example of significance testing, the process is as follows: Firstly, a test hypothesis is established and the threshold chosen for statistical significance

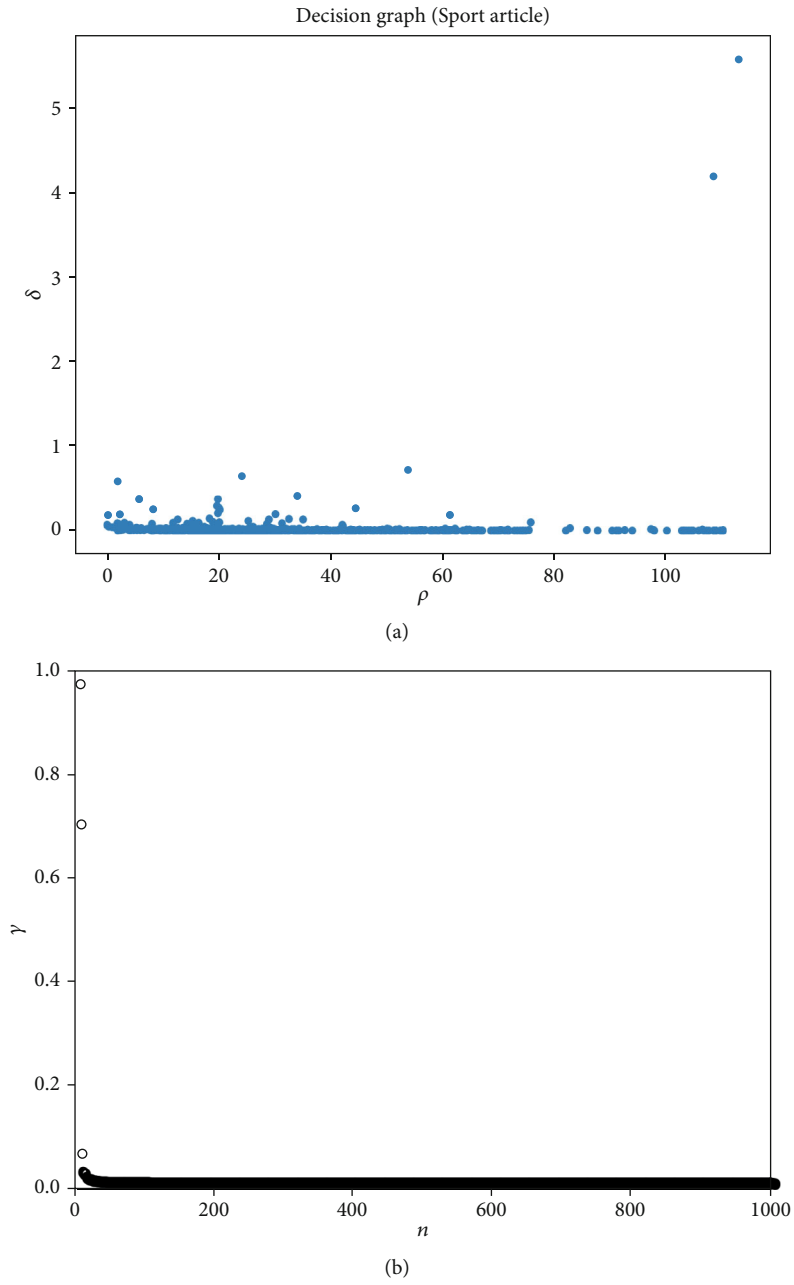

Figure 10: The improved DPC clustering of the Sports Article dataset. (a) Decision graph. (b) γ value.

was determined ($H_0 : \mu = \mu_0,\ \alpha = 0.1$). Secondly, the t is calculated:

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n} = \frac{0.9225 - \mu_0}{s}\sqrt{5} = \frac{0.9225 - 0.7362}{0.2657}\sqrt{5} = 1.568, \quad (11)$$
$$v = 5 - 1 = 4,$$

where $\bar{x}$ represents the FMI value obtained by DPC-K-means on the BBC dataset, μ_0 is the mean FMI of the five comparison algorithms on the BBC dataset, s represents the standard deviation FMI of the five comparison algorithms on the BBC dataset, and n is the sample size. The degree of freedom v used in this test is 4. Finally, the table was queried of t-distribution, the p value was determined, and an inference conclusion was made. According to $\alpha = 0.1$ and $v = 4$, the p value is 1.533, $t = 1.568 > p$, and H_0 is rejected, indicating the difference between the FMI metric of DPC-K-means on the BBC dataset, and the FMI value of other comparison algorithms is statistically significant. According to the above significance test steps, the t-test results of DPC-K-means were calculated on the ARI, NMI, and FMI evaluation metrics. The results are shown in Tables 11–13.

As shown in Table 11, there are significant differences in the ARI metric of DPC-K-means on five datasets, and as

Table 6: The ARI of each clustering algorithm.

Dataset		DPC-K-means	DPC	DBSCAN	Spectral Clustering	Affinity Propagation
BBC		0.9028	0.9002	0.4651	0.8961	0.1477
20-newsgroups	4 groups	0.9783	0.9781	0.7266	0.9756	0.1492
	5 groups	0.8521	0.8433	0.5993	0.8415	0.1508
	6 groups	0.6721	0.6759	0.4166	0.5130	0.1334
	7 groups	0.6078	0.5858	0.4260	0.4914	0.1480
	8 groups	0.4858	0.4664	0.1389	0.4599	0.1589
Sports Article		0.1941	0	0.0354	0.1906	0.0175
Asian Religious		0.1566	0.0189	0	0.1829	0.1064
Stack Overflow		0	0	0.0386	0	0.0349
Amazon		0	0.0014	0	0	0.0114

Table 7: The NMI of each clustering algorithm.

Dataset		DPC-K-means	DPC	DBSCAN	Spectral Clustering	Affinity Propagation
BBC		0.9028	0.8681	0.5652	0.8650	0.3696
20-newsgroups	4 groups	0.9763	0.9623	0.6725	0.9577	0.3628
	5 groups	0.8421	0.8381	0.6709	0.8404	0.3653
	6 groups	0.7721	0.7351	0.5721	0.6987	0.3304
	7 groups	0.7078	0.6487	0.5511	0.6498	0.3602
	8 groups	0.6858	0.5507	0.2846	0.6249	0.3719
Sports Article		0.1870	0	0.1286	0.1849	0.0307
Asian Religious		0.2673	0.0163	0	0.2443	0.2094
Stack Overflow		0	0	0.0518	0.0204	0.0819
Amazon		0	0.0048	0	0	0.0116

Table 8: The FMI of each clustering algorithm.

Dataset		DPC-K-means	DPC	DBSCAN	Spectral Clustering	Affinity Propagation
BBC		0.9225	0.9204	0.5805	0.9172	0.3402
20-newsgroups	4 groups	0.9823	0.9836	0.7930	0.9817	0.3225
	5 groups	0.8864	0.8846	0.6810	0.8831	0.2992
	6 groups	0.8486	0.7700	0.5485	0.6332	0.2593
	7 groups	0.7823	0.7031	0.5415	0.5813	0.2560
	8 groups	0.6535	0.6138	0.3896	0.5468	0.2592
Sports Article		0.7300	0.7298	0.5765	0.6114	0.1653
Asian Religious		0.4802	0.4665	0.4615	0.3833	0.2341
Stack Overflow		0.5512	0.4399	0.4004	0.3816	0.1649
Amazon		0.7192	0.6696	0.7041	0.6892	0.2624

shown in Table 12 there are significant differences in the NMI metric of DPC-K-means on eight datasets. It can be seen from Table 13 that DPC-K-means have significant difference in the FMI metric on the seven datasets. DPC-K-means are statistically significant on most datasets. Combined with Tables 9 and 10, it further shows that DPC-K-means is better than other comparison algorithms.

The clustering performance of DPC-K-means is better than K-means because DPC-K-means can select the number of clusters and obtain the initial clustering center. The

Table 9: The MSE of each clustering algorithm.

Dataset		DPC-K-means	DPC	DBSCAN	Spectral Clustering	Affinity Propagation
BBC		1.0661	3.279	13.2085	6.2378	15.4328
20-newsgroups	4 groups	0.5590	3.2087	6.2040	2.7103	14.6743
	5 groups	1.7203	2.2610	5.8610	3.0047	13.4050
	6 groups	6.6390	8.4530	6.9040	5.0280	13.7290
	7 groups	7.8453	8.0503	9.6103	7.2420	15.9880
	8 groups	5.4723	13.0367	16.6757	7.5143	20.1617
Sports Article		1.3020	2.0950	2.0980	1.8150	8.4410
Asian Religious		5.1898	6.4644	16.4831	11.9678	56.6118
Stack Overflow		3.8084	6.7365	18.0778	11.1916	60.4068
Amazon		0.4700	0.5200	0.5100	0.5300	5.2500

Table 10: Paired t-test results of clustering algorithms.

Pairing method	Paired t-test index	ARI	NMI	FMI	MSE
DPC-K-means and DPC	t	1.73	2.55	2.80	-2.88
	p	0.1171	0.0311	0.0207	0.0181
DPC-K-means and DBSCAN	t	4.39	3.93	5.47	-3.50
	p	0.0017	0.0035	0.0004	0.0067
DPC-K-means and Spectral Clustering	t	1.60	2.60	2.62	-2.35
	p	0.1448	0.0289	0.0056	0.0430
DPC-K-means and Affinity Propagation	t	3.68	3.72	12.52	-3.19
	p	0.0050	0.0047	0.0000	0.0109

Table 11: The results of the t-test of DPC-K-means on ARI.

Dataset		Mean	Standard deviation	t	p	Difference
BBC		0.6624	0.3438	1.398	1.533	No
20-newsgroups	4 groups	0.7617	0.3593	1.348	1.533	No
	5 groups	0.6574	0.3026	1.439	1.533	No
	6 groups	0.4822	0.2293	1.897	1.533	Yes
	7 groups	0.4518	0.1849	1.886	1.533	Yes
	8 groups	0.3420	0.1767	1.820	1.533	Yes
Sports Article		0.0875	0.0965	2.469	1.533	Yes
Asian Religious		0.093	0.0813	1.750	1.533	Yes
Stack Overflow		0.0842	0.1694	1.111	1.533	No
Amazon		0.0026	0.0050	1.150	1.533	No

number of clusters and the initial clustering centers can be used in the K-means algorithm, which achieves better clustering performance than K-means. Figures 11 and 12 illustrate the clustering centers automatically determined by DPC-K-means which are closer to the real class centers.

Tables 6–9 show that the clustering metrics changed significantly from 4 newsgroups to 8 newsgroups; this was caused by the loss of clustering due to irregular data distribution. Due to the inherent nature of the DPC algorithm, it cannot identify the phenomenon of "False peaks," and its clustering effect on "No density peaks" datasets is low, which are all factors that affect the accuracy of the DPC-K-means algorithm. The algorithm is limited in its processing of more complex datasets.

DPC-K-means has a parameter $\cos_c$, which is the cut-off distance. The value suggested in the literature [12] is set to

Table 12: The results of the t-test of DPC-K-means on NMI.

Dataset		Mean	Standard deviation	t	p	Difference
BBC		0.7141	0.2361	1.787	1.533	Yes
20-newsgroups	4 groups	0.7863	0.2687	1.581	1.533	Yes
	5 groups	0.7114	0.2069	1.413	1.533	No
	6 groups	0.6217	0.1794	1.875	1.533	Yes
	7 groups	0.5835	0.1369	2.029	1.533	Yes
	8 groups	0.5036	0.1699	2.399	1.533	Yes
Sports Article		0.1062	0.0869	2.078	1.533	Yes
Asian Religious		0.1475	0.1290	2.078	1.533	Yes
Stack Overflow		0.0308	0.0356	1.938	1.533	Yes
Amazon		0.0033	0.0051	1.440	1.533	No

Table 13: The results of the t-test of DPC-K-means on FMI.

Dataset		Mean	Standard deviation	t	p	Difference
BBC		0.7362	0.2657	1.568	1.533	Yes
20-newsgroups	4 groups	0.8126	0.2860	1.327	1.533	No
	5 groups	0.7269	0.2548	1.400	1.533	No
	6 groups	0.6119	0.2290	2.311	1.533	Yes
	7 groups	0.5728	0.2014	2.325	1.533	Yes
	8 groups	0.4926	0.1648	2.184	1.533	Yes
Sports Article		0.5626	0.2326	1.609	1.533	Yes
Asian Religious		0.4051	0.1028	1.632	1.533	Yes
Stack Overflow		0.3876	0.1408	2.598	1.533	Yes
Amazon		0.6089	0.1946	1.268	1.533	No

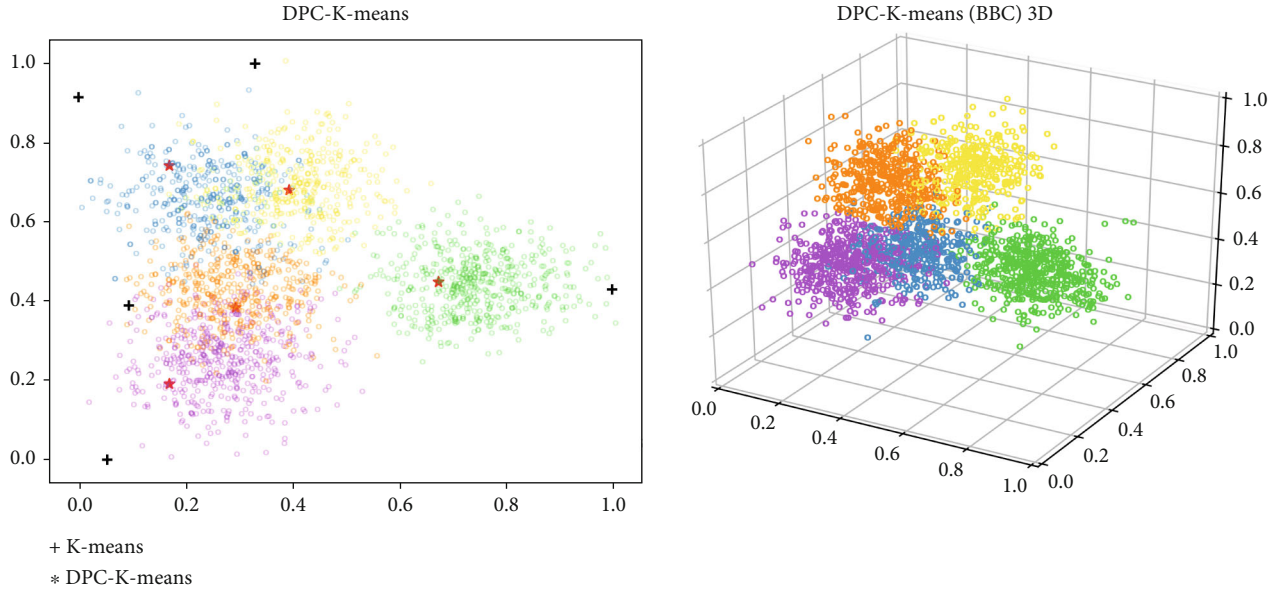

Figure 11: The BBC news dataset clustering center of K-means marked in black, and DPC-K-means clustering center marked in red and the 3D clustering results.

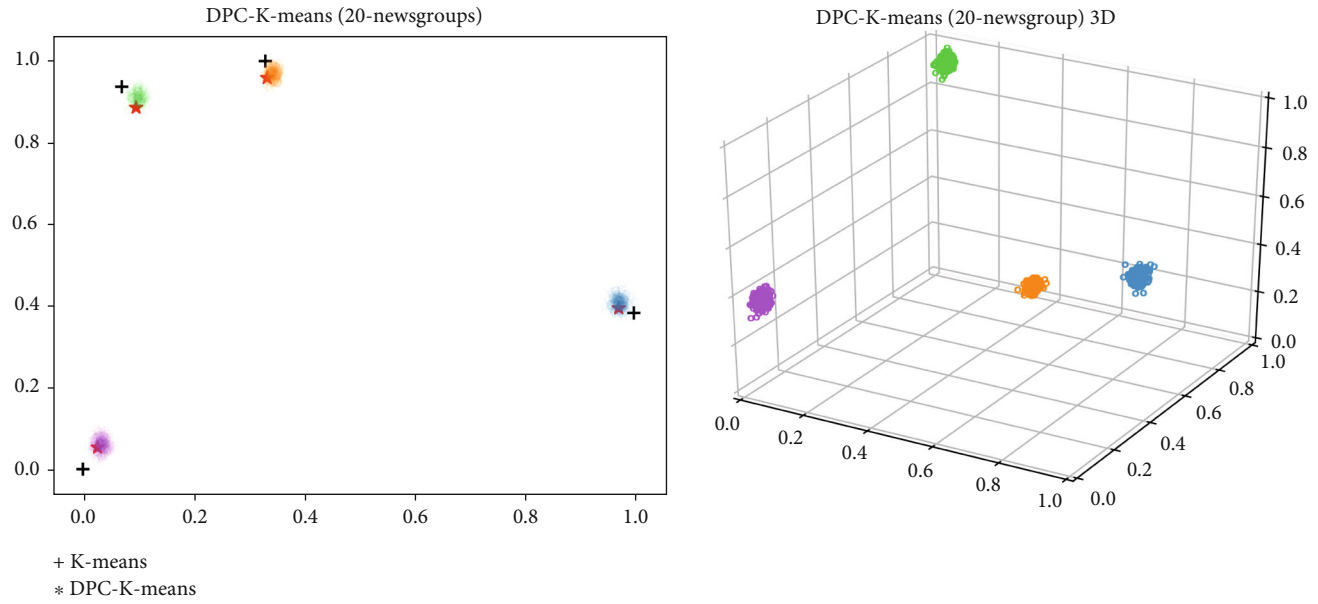

Figure 12: The 4 groups in the 20-newsgroups dataset clustering center of K-means marked in black, and DPC-K-means clustering center marked in red and the 3D clustering results.

the nearest neighbour number of the sample, approximately 1%~2% of the total dataset size. In the experiment, we were able to obtain the correct number of class clusters according to this valued principle. This parameter has no significant influence on the result of the algorithm within the value range of 1%–2% of the entire dataset size. The K-nearest neighbour method was used to establish the similarity matrix in the Spectral Clustering parameters in the experiment. The damping factor in Affinity Propagation was set to 0.9, and the nearest distance measurement of the DBSCAN parameter value was set to "cosine" by cosine similarity.

4. Conclusions

This study proposed a Stacked-Random Projection (SRP) dimension reduction framework based on deep networks and an improved K-means text clustering algorithm based on density peak (DPC-K-means). In the experiment, SRP, the improved DPC, and DPC-K-means were validated by using different datasets. Firstly, we compared SRP with PCA, MDS, and Random Projection. Multiple evaluation metrics demonstrated that SRP maintained a sufficient balance between running time and distance before and after dimension reduction. Secondly, we compared the difference between the Euclidean distance and cosine similarity in calculating DPC local density. Cosine similarity is more suitable for text vector calculation. Finally, DPC-K-means are an improved K-means algorithm that uses a text feature vector's cosine similarity to calculate local density and get the initial clustering center and cluster number. Then, the K-means algorithm is used for clustering. We compared DPC-K-means with DPC, DBSCAN, Spectral Clustering, and Affinity Propagation. We found that DPC-K-means can accurately determine the number of clusters and the initial clustering centers of high-dimensional text data. It is superior to other clustering algorithms in ARI, NMI, FMI, and MSE. Furthermore, we analyzed the influence of parameters on the algorithm and limitations of our proposed methods. We will focus on determining the number of layers and the target dimension of each layer dimensionality reduction for future work and improve the matching degree between DPC-K-means and datasets.

Data Availability

The BBC news data used to support the findings of this study have been deposited in the open-source repository (http://mlg.ucd.ie/datasets/bbc.html). The 20-newsgroups data used to support the findings of this study have been deposited in the open-source repository (http://scikit-learn.org/stable/modules/classes.html#module-sklearn.datasets).

Conflicts of Interest

Yujia Sun and Jan Platoš declare that there is no conflict of interest regarding the publication of this paper.

References

[1] C. Aggarwal and C. Zhai, *Mining text data*, Springer, New York, NY, 2012.

[2] T. Joachims, *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*, Norwell, MA, USA, 2002.

[3] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton university press, United States of America, 2015.

[4] X. S. Lu, M. C. Zhou, L. Qi, and H. Liu, "Clustering-algorithm-based rare-event evolution analysis via social media data," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 2, pp. 301–310, 2019.

[5] S. Zhou, X. Xu, Y. Liu, R. Chang, and Y. Xiao, "Text similarity measurement of semantic cognition based on word vector distance decentralization with clustering analysis," *IEEE Access*, vol. 7, pp. 107247–107258, 2019.

[6] A. Onan, "Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering," *IEEE Access*, vol. 7, pp. 145614–145633, 2019.

[7] J. Jokinen, T. Raty, and T. Lintonen, "Clustering structure analysis in time-series data with density-based clusterability measure," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 6, pp. 1332–1343, 2019.

[8] X. Xu, J. Li, M. C. Zhou, J. Xu, and J. Cao, "Accelerated two-stage particle swarm optimization for clustering not-well-separated data," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 11, pp. 4212–4223, 2020.

[9] L. Liu, A. Yang, W. Zhou, X. Zhang, M. Fei, and X. Tu, "Robust dataset classification approach based on neighbor searching and kernel fuzzy c-means," *IEEE/CAA Journal of Automatica Sinica*, vol. 2, no. 3, pp. 235–247, 2015.

[10] K. Orkphol and W. Yang, "Sentiment analysis on microblogging with K-means clustering and artificial bee colony," *International Journal of Computational Intelligence and Applications*, vol. 18, no. 3, p. 1950017, 2019.

[11] U. H. Atasever, "A novel unsupervised change detection approach based on reconstruction independent component analysis and ABC-Kmeans clustering for environmental monitoring," *Environmental Monitoring and Assessment*, vol. 191, no. 7, 2019.

[12] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[13] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189–206, 1984.

[14] J. Zhang, M. Zhu, P. Chen, and B. Wang, "DrugRPE: random projection ensemble approach to drug-target interaction prediction," *Neurocomputing*, vol. 228, pp. 256–262, 2017.

[15] L. Gondara, "RPC: an efficient classifier ensemble using random projection," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pp. 559–564, Miami, FL, USA, December 2015.

[16] S. Sieranoja and P. Fränti, "Fast and general density peaks clustering," *Pattern Recognition Letters*, vol. 128, pp. 551–558, 2019.

[17] M. Parmar, D. Wang, X. Zhang et al., "REDPC: a residual error-based density peak clustering algorithm," *Neurocomputing*, vol. 348, pp. 82–96, 2019.

[18] M. D. Parmar, W. Pang, D. Hao et al., "FREDPC: a feasible residual error-based density peak clustering algorithm with the fragment merging strategy," *IEEE Access*, vol. 7, pp. 89789–89804, 2019.

[19] M. Parmar, D. Wang, A. Tan, C. Miao, J. Jiang, and Y. Zhou, "A novel density peak clustering algorithm based on squared residual error," in *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pp. 43–48, Shenzhen, China, December 2017.

[20] D. Cheng, Q. Zhu, J. Huang, Q. Wu, and L. Yang, "A novel cluster validity index based on local cores," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 4, pp. 985–999, 2019.

[21] D. Cheng, Q. Zhu, J. Huang, Q. Wu, and Y. Lijun, "Clustering with local density peaks-based minimum spanning tree," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, p. 1, 2019.

[22] F. Heimerl, M. John, Q. Han, S. Koch, and T. Ertl, "DocuCompass: effective exploration of document landscapes," in *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 11–20, Baltimore, MD, USA, Oct 2016.

[23] B. Wang, J. Zhang, F. Ding, and Y. Zou, "Multi-document news summarization via paragraph embedding and density peak clustering," in *2017 International Conference on Asian Language Processing (IALP)*, pp. 260–263, Yuexian Zou, Dec 2017.

[24] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.

[25] P. Krömer and J. Platoš, "Cluster analysis of data with reduced dimensionality: an empirical study," in *Intelligent Systems for Computer Modelling*, pp. 121–132, Springer, Cham, 2016.

[26] P. Fränti and S. Sieranoja, "How much can k-means be improved by using better initialization and repeats?," *Pattern Recognition*, vol. 93, pp. 95–112, 2019.

[27] T. Sung, L. Kong, P. Tsai, and J. Pan, "A distance coefficient-based algorithm for k-center selection in wireless sensor networks," in *2017 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pp. 293-294, Taipei, Taiwan, June 2017.