An Improved Numerical DBSCAN Algorithm Based on Non-IIDness Learning

YANMEI WANG, HE JIANG, HUIXIN ZHOU, AND CHENGQIANG WANG
School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
Corresponding author: He Jiang (jianghe09@126.com)

This work was supported by the National Natural Science Foundation of China under Grant 61502259.

ABSTRACT In clustering algorithm research, objects, attributes and other aspects of data sets are usually considered to be independent and identically distributed; that is, each object is assumed to be an independent and uniformly distributed individual with no impacts between objects. However, objects in real life are often neither independently nor identically distributed; that is, they are non-IID, leading to a complex coupling relationship between objects, and objects interact with each other. The results of a clustering algorithm under an independent and identical distribution may be incomplete or even misleading. To make the results of the DBSCAN algorithm as accurate as possible, an improved numerical DBSCAN algorithm based on non-IIDness learning is proposed in this paper. The algorithm calculates the coupling relationship between objects to obtain the potential relationship between objects and determines the parameters $\epsilon$ and $MinPts$ by the distribution characteristics of the data. Experiments on large-scale real and synthetic data sets show that the algorithm achieves a higher accuracy than the original DBSCAN algorithm and the main algorithms that improved upon it.

INDEX TERMS Clustering, DBSCAN, parameter selection, non-IIDness learning.

I. INTRODUCTION

Clustering refers to the grouping of abstract or physical objects in accordance with the principle that objects in groups should be as similar to each other as possible and that objects in different groups be as different as possible under the condition that samples are not marked; the ultimate purpose of clustering is to discover the natural structure of data [1]. Clustering is an unsupervised machine learning method and therefore has its own unique advantages to deal with some data sets with unknown distributions. Clustering has become an important component of machine learning, pattern recognition, data mining and other fields.

Over many years of research and development, clustering algorithms have been greatly developed; there are many types of clustering algorithms that, according to different clustering principles and ideas, can be divided into several types of clustering algorithms based on models, divisions, hierarchies, densities, grids, and so on. Clustering techniques are diverse, and different data show different characteristics. Therefore, different types of data have different requirements for clustering techniques. For numerical data, the model-based clustering algorithm usually uses a mixed Gao Si model, represented as an EM algorithm [2]. Clustering algorithms based on partitioning divide the data set into several groups according to a certain target, and the result is to maximize or minimize the value of the objective function; for numerical data, clustering algorithms based on partitioning are represented by k-means [3], [4], k-medoids [5], and k-medians [6]. Hierarchical clustering algorithms aim to form a tree-shaped clustering structure, often using a “ottom-up” aggregation strategy or a “top-down” splitting strategy to divide data sets; for numerical data, HC [7], CURE [8], and BRICH [9] hierarchical representation algorithms are used. Density-based clustering algorithms interpret the density of data points as the number of data points contained in designated neighbourhoods; such algorithms consider the connectivity between samples according to the density of samples and continuously expand the cluster based on connected samples to obtain clustering results. Density clustering is a method to determine the clustering structure by the density of the sample distribution. Classic density-based clustering methods for numerical data are DBSCAN [10], OPTICS [11], and DENCLUE [12]. Grid-based clustering can be regarded as the implementation of a density-based clustering algorithm, which takes the number of data points in each grid as the density and spices...
adjacent dense grids to form clusters. Grid-based representative clustering algorithms for numerical data include GRIDCLUS [13], BANG [14], and STING [15].

DBSCAN (density-based spatial clustering of applications with noise) is a classic density-based clustering algorithm. Its advantage is that it can not only find clusters of arbitrary shapes and sizes in data but also find noise points and outliers in data. To date, DBSCAN algorithms have been applied to consumer market analysis [16], vehicle identification [17], time series processing [18], blending images [19], detection of abnormal population behavior [20], abnormal flight detection [21] and so on.

Although the DBSCAN algorithm has great advantages and extensive applications, the accuracy is not high enough for certain applications, which limits the use of the DBSCAN algorithm to a certain extent. Therefore, it is necessary to further improve the accuracy of the algorithm. To improve the accuracy of DBSCAN clustering results, scholars worldwide have improved and innovated the algorithm in different directions. Greedy-DBSCAN algorithm [22] adaptively determines the local optimal \( \text{EPS} \) and detects noise based on the \( \text{EPS} \) value. However, the algorithm has difficulty determining the threshold parameters required for noise determination. AD-DBSCAN algorithm [23] adaptively determines the parameters of the algorithm, reduces the error caused by manual parameter settings, and thus improves the accuracy of the clustering results of the algorithm. However, the algorithm cannot automatically identify the number of cluster classes, and thus, the number of clusters needs to be specified in advance. A new method [24] to calculate the \( \text{EPS} \) and \( \text{MinPts} \) parameters of the DBSCAN algorithm improved the parameter selection accuracy. However, if the knee is not accurately specified in the sorting distance, it may cause incorrect clustering results. FDBSCAN algorithm [25] improved the calculation speed by reducing the distance measurement when searching the core object while ensuring the clustering accuracy. However, when the size of the data set was small, its clustering accuracy was low. 3W-DBSCAN algorithm [26] represents a cluster through a pair of nested sets called the lower bound and end segment, which could explain the clustering results and was in line with human cognitive thinking. However, for some data sets, the clustering results of the 3W-DBSCAN algorithm overlapped. AA-DBSCAN algorithm [27] uses a new tree structure based on a quadtree to define the data set density layer. This method allows AFM to find clusters of different densities more accurately. However, the use of inflexible \( \varepsilon \text{DL} \) can lead to incorrect clustering results. KLS-DBSCAN algorithm [28] automatically selects the globally optimal \( \text{EPS} \) and \( \text{MinPts} \) parameters by a mathematical statistics method, establishing a mathematical model. Although KLS-DBSCAN can realize the adaptive selection of parameters, data clustering classes with a large density difference have a general effect. Non-IIDDBSCAN algorithm [29] aiming at the problem that the traditional distance formula cannot accurately calculate the similarity of categorical data in non-independent co-distributed environments, the non-IIDDBSCAN algorithm replaces the traditional distance formula with the coupling similarity. The accuracy of the algorithm is improved, but the non-IIDDBSCAN algorithm is only applicable to categorical data.

At present, in addition to the non-IIDDBSCAN algorithm, other improvements to the accuracy of the DBSCAN algorithm are based on independent identical distributions. This assumption ignores the internal relation between data points. In reality, data are often neither independently nor identically distributed [30]. Each data in the data set has more or less interaction, that is, each data has non-IID property. And different types of data have different characteristics of interaction relationships. Therefore, for numerical data, we propose the NDBSCAN algorithm, which is based on the improvement of the DBSCAN algorithm in the case of a non-independent identical distributions. This algorithm focuses on the non-IID characteristics of each data in the data set and uses the principle of coupling similarity to quantify the relationship between the data. Then, the parameter neighborhood \( \text{EPS} \) and density threshold \( \text{MinPts} \) are selected by using the distribution of data. Experimental results show that the clustering results of this algorithm are more accurate than those of the comparison algorithms.

II. THE THEORY OF DBSCAN CLUSTERING AND NUMERICAL DATA COUPLING

ATTRIBUTE ANALYSIS

A. DBSCAN CLUSTERING THEORY

DBSCAN is a classic density-based clustering algorithm that seeks the final result of clustering by finding the maximum set of connected data point densities.

For data set \( D = \{x_1, x_2, \ldots, x_n\} \), the definitions are as follows:

Definition 1 (Neighborhood \( \text{EPS} \)) [10]: The neighborhood of \( x_i \) is the supersphere region with the centre of \( x_i \) and the radius of the \( \text{EPS} \). In addition, \( N_{\text{Eps}}(x_i) = \{x_j \in D | \text{dist}(x_i, x_j) \leq \text{Eps}\} \), where \( x_i \) is a set of sample points and \( \text{dist}(x_i, x_j) \) represents a measure of distance, for example, the Euclidean distance between \( x_i \) and \( x_j \).

Definition 2 (Density Threshold MinPts) [10]: If for \( x_i \in D \), \( \text{MinPts} \) is the minimum density that makes \( x_i \) become the core point.

Definition 3 (Core and Boundary Points) [10]: \( x_i \) is a core point if the \( \text{EPS} \) value of \( x_i \) is greater than or equal to the \( \text{MinPts} \) objects. If \( |N_{\text{Eps}}(x_i)| \geq \text{MinPts} \), then \( x_i \) is a core point. If a point is a non-nuclear point but is in the \( \text{EPS} \) neighbourhood of a core point, the point is called a boundary point.

Definition 4 (Directly Density-Reachable) [10]: If \( x_i \) is the core object and \( x_j \) is in the \( \text{EPS} \) neighbourhood of \( x_i \), then \( x_i \) does not satisfy symmetry and is said to be directly identically reachable with respect to \( \text{EPS} \) and \( \text{MinPts} \) of \( x_j \).

Definition 5 (Density-Reachable) [10]: If there exists a chain \( P_1, P_2, \ldots, P_n \), if \( P_1 = x_i \), \( P_n = x_j \), and \( P_{i+1} \) is directly density-reachable from \( P_i \), then \( x_i \) is the density that \( x_j \) can reach.
Definition 6 (Density-Connected) [10]: For \( x_i \) and \( x_j \), if there is \( x_k \) so that both \( x_i \) and \( x_j \) are reachable by the density of \( x_k \), then \( x_i \) and \( x_j \) are density-connected, and the density connection has symmetry.

Definition 7 (Cluster and Noise Points) [10]: Any core point, all objects with an object density up to form a cluster, and points that do not belong to any cluster are marked as noise points.

### B. THE THEORY OF NUMERICAL DATA COUPLING ATTRIBUTE ANALYSIS

Cao put forward the idea of non-IIDness in 2014, which proposes that data do not exist independently but have more or less explicit or implicit relationships with other data. Wang proposed a coupling representation scheme [31] of numerical objects by integrating the internal coupling of continuous attributes and the coupling [32] interaction between attributes on the basis of the idea of non-IIDness. For a numerical data set \( D \), the coupling attribute analysis framework is shown in Fig. 1.

### C. REASONS TO IMPROVE THE COUPLING PROCESS IN DBSCAN

The clustering algorithm is a process to derive the sample set with the maximum density relation from the density-reachable relation. A process diagram of clustering a data set using the DBSCAN algorithm is shown in Fig. 2. In the figure, \( \text{EPS} = d \), \( \text{MinPts} = 4 \), red dots are core objects because there are at least four samples in the \( \text{EPS} \) neighbourhood, and black dots are non-core objects. All core object density direct samples in the red core objects are in the centre of the superball body. If they are not in the superball body, they cannot be density direct. The core objects connected by green arrows form a density-reachable sample sequence, and the set of samples connected by the maximum density derived from the density-reachable relation is a category of final clustering. Furthermore, all samples in the \( \text{EPS} \) neighbourhood of these density-accessible sample sequences are connected to each other.

In Fig 2, each point is an object. The NDBSCAN algorithm proposed in this paper needs to first calculate the relationship between each point to obtain the coupling representation of the original data. This process will aggravate the computation of the DBSCAN algorithm. Therefore, the coupling calculation of the data set is improved to reduce the computation of the DBSCAN algorithm.

### D. THE OPTIMIZED COUPLING ATTRIBUTE ANALYSIS OF NUMERICAL DATA

For data set \( D \), the improved coupling calculation process is shown below.

\[
D = \begin{bmatrix}
  a_{11} & a_{12} & \cdots & a_{1m} \\
  a_{21} & a_{22} & \cdots & a_{2m} \\
  \vdots & \vdots & \ddots & \vdots \\
  a_{n1} & a_{n2} & \cdots & a_{nm}
\end{bmatrix}
\]

where \( a_{nm} \) is the \( m \)th attribute value of the \( n \)th object and the value of each attribute in data set \( D \) is extended to a power of 2. The extended data set is generated as \( D_1 \):

\[
D_1 = \begin{bmatrix}
  a_{11}^1 & a_{11}^2 & \cdots & a_{1m}^1 & a_{1m}^2 \\
  a_{21}^1 & a_{21}^2 & \cdots & a_{2m}^1 & a_{2m}^2 \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  a_{n1}^1 & a_{n1}^2 & \cdots & a_{nm}^1 & a_{nm}^2
\end{bmatrix}
\]
where \( a_{m}^2 \) is the second power value of the \( m \)th attribute of the \( n \)th object. To facilitate writing the following formula, \( a_{nm} \) is written as \( a_m \), and \( a_{nm}^2 \) is written as \( a_m^2 \).

Pearson correlation coefficients [33] are used to quantify the interaction between attribute values. The Pearson coefficient between attributes \( a_i \) and \( a_j \) is formalized as shown in 1.

\[
Cor(a_i, a_j) = \frac{\sum_{u \in U} u(U_i - \mu_i)(U_j - \mu_j)}{\sqrt{\sum_{u \in U}(U_i - \mu_i)^2} \sqrt{\sum_{u \in U}(U_j - \mu_j)^2}}
\]  

where \( f_i(u) \) is the value of object \( u \) on attribute \( a_i \), \( f_j(u) \) is the attribute value of object \( u \) on attribute \( a_j \), and \( \mu_i \) and \( \mu_j \) are the average values of \( a_i \) and \( a_j \), respectively.

When quantifying the interaction relationship between attributes, this paper only considers the important attribute coupling relationship rather than simply involving all attributes to avoid overfitting the coupling relationship. Therefore, 0.05 is used as the dividing point of the P value, and the modified formula is shown in Formula 2.

\[
Cor(a_i, a_j) = \begin{cases} 
Cor(a_i, a_j) & \text{if } p - value < 0.05, \\
0 & \text{if } p - value \geq 0.05.
\end{cases}
\]  

According to the revised Pearson correlation coefficient, the internal coupling and the inter-attribute coupling of the attributes are calculated separately.

The internal coupling is quantified as the correlation between attribute \( a_i \) and its quadratic power attribute value, which is expressed as \( R^{a_i}(a_i) \).

\[
R^{a_i}(a_i) = \begin{bmatrix} \varphi_{11}(i) & \varphi_{12}(i) \\
\varphi_{21}(i) & \varphi_{22}(i) \end{bmatrix}
\]  

where \( \varphi_{ii}(i) = \varphi_{21}(i), \varphi_{jk}(i) = Cor(d_i, d_j^k) \), and \( d_i, d_j \) represent the \( j \) and \( k \) powers of the \( a_i \) attribute, respectively.

Coupling between attributes is quantified as the correlation between the \( a_i \) and \( a_j \) attributes, which is expressed as \( \text{Re}^{a_i}(a_i | [a_j]_{j \neq i}) \).

\[
\text{Re}^{a_i}(a_i | [a_j]_{j \neq i}) = \begin{bmatrix} e_{11}(i) & e_{12}(i) & \cdots & e_{11}(i) & e_{12}(i) & e_{12}(i) \\
e_{21}(i) & e_{22}(i) & \cdots & e_{21}(i) & e_{22}(i) & e_{22}(i) \\
e_{21}(n-1) & e_{22}(n-1) & \cdots & e_{21}(n-1) & e_{22}(n-1) & e_{22}(n-1) \\
e_{21}(n-2) & e_{22}(n-2) & \cdots & e_{21}(n-2) & e_{22}(n-2) & e_{22}(n-2) \\
e_{21}(1) & e_{22}(1) & \cdots & e_{21}(1) & e_{22}(1) & e_{22}(1) \end{bmatrix}
\]  

In 4, \( e_{pq}(i) = Cor(d_i^p, d_j^q) \). The coupling representation of numerical objects is obtained by integrating the internal coupling and inter-attribute coupling of continuous attributes, as shown in Formula 5.

\[
U_n(a_i | D_i) = U_n(a_i) \odot \beta \odot R^{a_i}(a_i)
\]

where \( U_n(a_i) \) is the original value of the \( i \)th attribute of the \( n \)th object and its second power value. \( U_n(a_i) \) refers to all values of the \( n \)th object except \( U_n(a_i) \). \( \odot \) represents the Hadamard product, \( \odot \) represents the matrix product, \( \beta = [1/(1!), 1/(2!), \ldots, 1/(L!)] \), and \( [\beta, \beta, \ldots, \beta] \) represents a vector with \( (n-1) \beta \) elements.

### III. PROPOSED ALGORITHM

To cluster numerical data effectively, we propose an improved DBSCAN algorithm based on the idea of a non-independent and identical distribution, namely, the NDBSCAN algorithm. The algorithm first quantifies the relationship between data by finding the coupling relationship between data points and then selects the \( \text{EPS} \) and \( \text{MinPts} \) parameters according to the distribution of data. The flowchart of the NDBSCAN algorithm is shown in Fig. 3.

**FIGURE 3.** The flowchart of the NDBSCAN clustering algorithm.

### A. COUPLING ATTRIBUTE ANALYSIS

The improved coupling analysis process in Section II is used to analyse and calculate the numerical data set to obtain the coupling data set that can reflect the relationship between the data.

### B. SELECTING THE EPS PARAMETER

First, the distance between objects in the coupled data set needs to be calculated. The distance formula is shown in Formula 6, and the distance distribution matrix \( \text{DIST}_{n \times n} \) is obtained as shown in Formula 7.

\[
dist(X, Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}
\]

\[
\text{DIST}_{n \times n} = \{\text{dist}(X, Y)|1 \leq X, Y \leq n\}
\]

where \( X \) and \( Y \) are objects in the coupled data set and the \( \text{DIST}_{n \times n} \) matrix is a real symmetric matrix with zeros along the diagonal. The elements in each column of \( \text{DIST}_{n \times n} \) are sorted in increasing order, and the element in row \( m \) is the distance from each object to the nearest neighbour object in row \( m \) (m-1). After sorting, the k-distance is plotted, as shown in Fig. 4. In Fig. 4, lines of different colours represent different objects. Each line shows the Euclidean distance between it and other objects, and the Euclidean distance is arranged in increasing order.

In the range of the densest distribution of KNN lines, that is, at the red box in Fig. 4, the ordinate of the point with the largest slope change is the \( \text{EPS} \) value.

### C. SELECTING THE MINPTS PARAMETER

\[
\text{MinPts} = \begin{cases} 
4, & x \leq 150 \\
n/25, & x > 150
\end{cases}
\]
IV. EXPERIMENTAL RESULTS AND ANALYSIS

The NDBSCAN algorithm is implemented in MATLAB and runs in the Windows 10 system. The processor used is an AMD Ryzen 5 4600U with a Radeon Graphics 2.10 GHz graphics card and 16 GB of memory.

To verify the effectiveness of the algorithm, experiments were carried out on real data sets and synthetic data sets, and the performance of the DBSCAN [10], F-DBSCAN [25], 3W-DBSCAN [26], AA-DBSCAN [27], and KLS-DBSCAN [28] algorithms was compared through experiments.

A. DATA SETS

We used four synthetic data sets downloaded from the website “Clusters Datasets” [34]: Flame, R15, Aggregation and S1. The other synthetic data sets were downloaded from GitHub [35]. Eleven real data sets were downloaded from the UCI machine learning repository [36] to determine the ability of the algorithm to process data sets of different densities. Table 1 shows the attributes of the data sets, including their different data sizes and dimensions.

B. EVALUATION INDICATORS

To reflect the quality of the final clustering results with data, this paper uses the unsupervised clustering accuracy [37], F-measure [38], ARI [39] and NMI [40] as the evaluation indexes of the algorithm. The following is a brief introduction of these two indexes.

1) ACCURACY

Accuracy is used to compare whether the obtained label is consistent with the real label of the data set, with a value range of [0,1]. The closer the value is to 1, the more consistent it is with the real clustering result, which is defined as:

\[
\text{Accuracy} = \frac{\sum_{i=1}^{N} \delta(s_i, \text{map}(r_i))}{N}
\]  

(9)

where \(s_i\) is the true label, \(r_i\) is the label after clustering, map represents the reproduction allocation of the best class index, \(N\) represents the total number of data sets, \(\delta\) is the indicator.
TABLE 1. Synthetic and real data sets used in the experiments.

| Type  | Dataset | Data size | Dimension | Clusters |
|-------|---------|-----------|-----------|----------|
| Synthetic | Moon   | 210       | 2         | 2        |
|        | Flame  | 240       | 2         | 2        |
|        | Smile  | 266       | 2         | 3        |
|        | 2circles | 600     | 2         | 2        |
|        | R15    | 600       | 2         | 15       |
|        | Aggregation | 788    | 2         | 7        |
|        | Size5  | 1000      | 2         | 4        |
|        | Square | 1000      | 2         | 4        |
|        | Three circles | 3603   | 2         | 3        |
|        | S1     | 5000      | 2         | 15       |
| Real   | Solar  | 72        | 18        | 2        |
|        | Iris   | 150       | 4         | 3        |
|        | Wine   | 178       | 13        | 3        |
|        | Sonar  | 208       | 60        | 2        |
|        | Seeds  | 210       | 7         | 3        |
|        | Ecoli  | 336       | 7         | 8        |
|        | Blood  | 748       | 4         | 2        |
|        | Banknote | 1372    | 4         | 2        |
|        | Magic  | 2500      | 10        | 2        |
|        | Phoneme | 5404    | 5         | 2        |
|        | Robotnavigation | 5456  | 24        | 4        |

function, and the specific formula of $\delta$ is as follows:

$$\delta(x, y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise} \end{cases}$$  \hspace{1cm} (10)$$

2) F-MEASURE

As a comprehensive measure of the accuracy and recall rate, the F-measure can effectively reflect the performance of the clustering algorithm. The value of the F-measure represents the clustering effect. The larger the value, the better the clustering effect. It is defined as:

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$  \hspace{1cm} (11)$$

where Recall is the recall rate. The proportion of objects in the predefined cluster correctly divided into clusters by the clustering algorithm is calculated. Precision is the ratio of the objects in the cluster found by the clustering algorithm to the correct cluster. The recall and precision calculation methods are as follows:

$$\text{Recall} = \sum \frac{n_{jk}}{n_j}$$  \hspace{1cm} (12)$$

$$\text{Precision} = \sum \frac{n_{jk}}{n_k}$$  \hspace{1cm} (13)$$

where $n_{jk}$ is the number of samples correctly divided into cluster $k$ in cluster $j$, $n_j$ represents the cluster generated by the clustering algorithm, and $n_k$ represents the actual correct cluster of the dataset.

FIGURE 5. Synthetic data sets. a, b, c, d, e, f, g, h, i and j are the Moon, Flame, Smile, 2circles, R15, Aggregation, Size5, Square, Three circles, and S1 data sets, respectively.

3) ARI

The ARI reflects the degree of overlap between the two classifications, and the use of this metric requires that the data
The NMI value ranges from 0 to 1. If the NMI value is close to 1, the clustering quality is high.

### C. EXPERIMENTAL RESULTS AND ANALYSIS

This paper discusses the clustering performance of each algorithm in a reasonable parameter search range. DBSCAN, FDBSCAN, and NDBSCAN use the global parameters \( EPS \) and \( MinPts \). The selection of parameters depends on the data set, and the clustering results are sensitive to both parameters. In this study, the parameters that make the F-measure value of each data set high are selected through multiple iterations. 3W-DBSCAN uses a pair of nested sets, called the lower

---

**TABLE 2. Parameters and their search ranges for each algorithm.**

| Algorithm       | Parameters and search ranges                                  |
|-----------------|----------------------------------------------------------------|
| DBSCAN          | \( \text{Eps} \in [0,0.5]; ~\text{MinPts} \in \{3,4,…,15\} \)   |
| F-DBSCAN        | \( \epsilon \in [0.15,0.25]; ~\text{MinPts} \in [6,10] \)      |
| AA-DBSCAN       | \( \text{MinPts} \in \{3,4,…,100\} \)                           |
| KLS-DBSCAN      | \( \text{Eps} \in [0,0.5]; ~\text{MinPts} \in \{3,4,…,15\} \)  |
| NDBSCAN         | \( \text{Eps} \in [0,1000000]; ~\text{MinPts} \in \{3,4,…,20\} \) |

---

Y. Wang et al.: Improved Numerical DBSCAN Algorithm Based on Non-IIDness Learning
and upper bounds, to describe each cluster. The union of the positive region and the boundary region of clustering constitutes the upper bound, while the positive region constitutes the lower bound. In this paper, the average value of the clustering evaluation indexes of the upper bound and the lower bound is used to represent the overall clustering effect of 3W-DBSCAN. AA-DBSCAN uses MinPts as one parameter. KLS-DBSCAN uses a mathematical statistical method.
TABLE 3. (Continued.) Comparison among the four benchmarks of the six clustering algorithms on the synthetic data sets.

|       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|
| R15   | F1    | 0.872 | 0.852 | 0.926 | 0.960 | 0.933 | **0.962** |
|       | Acc   | 0.850 | 0.849 | 0.899 | 0.953 | 0.910 | **0.955** |
|       | ARI   | 0.887 | 0.896 | 0.910 | **0.937** | 0.917 | 0.935 |
|       | NMI   | 0.834 | 0.815 | 0.897 | 0.911 | 0.898 | **0.927** |
| Aggregation | F1 | 0.956 | 0.901 | **0.979** | 0.921 | 0.916 | 0.935 |
|       | Acc   | 0.928 | 0.914 | 0.937 | 0.956 | 0.922 | **0.970** |
|       | ARI   | 0.937 | 0.914 | 0.939 | 0.948 | 0.914 | **0.969** |
|       | NMI   | 0.915 | 0.899 | 0.932 | 0.950 | 0.918 | **0.968** |
| Size5 | F1    | 0.945 | 0.906 | 0.956 | 0.947 | 0.942 | **0.963** |
|       | Acc   | 0.928 | 0.924 | 0.934 | 0.933 | 0.926 | **0.987** |
|       | ARI   | 0.935 | 0.921 | 0.929 | 0.938 | 0.923 | **0.940** |
|       | NMI   | 0.916 | 0.931 | 0.938 | 0.951 | 0.918 | **0.955** |
| Square| F1    | **0.965** | 0.915 | 0.930 | 0.945 | 0.967 | 0.963 |
|       | Acc   | 0.946 | 0.934 | 0.935 | 0.922 | 0.935 | **0.954** |
|       | ARI   | 0.945 | 0.927 | 0.929 | 0.934 | 0.942 | **0.949** |
|       | NMI   | 0.965 | 0.924 | 0.897 | 0.974 | 0.954 | **0.976** |
| Three | F1    | 0.924 | 0.925 | 0.929 | 0.919 | **0.935** | 0.930 |
to establish a mathematical model and automatically selects the global optimal $\text{EPS}$ and $\text{MinPts}$ parameters. Since the parameter setting of each algorithm is different according to each data set, this paper presents the parameters and search range of each algorithm except the 3W-DBSCAN algorithm, as shown in Table 2.

### TABLE 3. (Continued.) Comparison among the four benchmarks of the six clustering algorithms on the synthetic data sets.

| Algorithm | circles | Acc   | F1   | NMI  |
|-----------|---------|-------|------|------|
|           | 0.938   | 0.942 | 0.956| 0.954| 0.953|
| ARI       | 0.926   | 0.939 | 0.936| 0.941| 0.947| 0.912|
| NMI       | 0.961   | 0.963 | 0.935| 0.936| 0.948| 0.967|
| S1        | 0.932   | 0.903 | 0.950| 0.850| 0.924| 0.971|
| F1        | 0.911   | 0.899 | 0.945| 0.891| 0.947| 0.978|
| Acc       | 0.900   | 0.906 | 0.935| 0.920| 0.945| 0.947|
| ARI       | 0.900   | 0.906 | 0.935| 0.920| 0.945| 0.947|
| NMI       | 0.897   | 0.897 | 0.902| 0.948| 0.921| 0.989|

### FIGURE 7. Four evaluation index values of ten real datasets on six clustering algorithms.
TABLE 4. Comparison among the four benchmarks for the six clustering algorithms on the real data sets.

| Dataset | Validity | Algorithm measure | DBSCAN | FDBSCAN | 3W-DBSCAN | AA-DBSCAN | KLS-DBSCAN | NDBSCAN |
|---------|----------|-------------------|--------|---------|-----------|-----------|------------|---------|
| Sobar   | F1       | 0.652             | 0.631  | 0.675   | 0.601     | 0.634     | **0.681**  |         |
|         | Acc      | 0.665             | 0.620  | 0.700   | 0.689     | 0.645     | **0.701**  |         |
|         | ARI      | 0.653             | 0.628  | 0.668   | 0.634     | 0.652     | **0.694**  |         |
|         | NMI      | 0.667             | 0.630  | 0.683   | 0.649     | 0.647     | **0.711**  |         |
| Iris    | F1       | 0.770             | 0.638  | 0.827   | **0.860** | 0.673     | 0.849      |         |
|         | Acc      | 0.767             | 0.649  | 0.720   | 0.812     | 0.687     | **0.827**  |         |
|         | ARI      | 0.754             | 0.646  | 0.765   | 0.823     | 0.550     | **0.854**  |         |
|         | NMI      | 0.769             | 0.639  | 0.848   | 0.819     | 0.669     | **0.850**  |         |
| Wine    | F1       | 0.568             | 0.812  | 0.612   | 0.650     | 0.659     | **0.680**  |         |
|         | Acc      | 0.621             | 0.578  | 0.600   | 0.649     | 0.633     | **0.655**  |         |
|         | ARI      | 0.617             | 0.624  | 0.619   | 0.637     | 0.642     | **0.663**  |         |
|         | NMI      | 0.609             | 0.631  | 0.621   | 0.642     | 0.637     | **0.659**  |         |
| Sonar   | F1       | 0.623             | 0.621  | 0.624   | 0.641     | **0.652** | 0.648      |         |
|         | Acc      | 0.628             | 0.627  | 0.615   | 0.639     | 0.634     | **0.646**  |         |
|         | ARI      | 0.621             | 0.630  | 0.623   | 0.617     | 0.629     | **0.639**  |         |
TABLE 4. (Continued.) Comparison among the four benchmarks for the six clustering algorithms on the real data sets.

|                  | NMI  | 0.619 | 0.609 | 0.630 | 0.628 | 0.633 | 0.647 |
|------------------|------|-------|-------|-------|-------|-------|-------|
| **Seeds**        |      |       |       |       |       |       |       |
| F1               | 0.607| 0.652 | 0.616 | 0.810 | 0.679 | 0.810 |
| Acc              | 0.656| 0.721 | 0.652 | 0.740 | 0.687 | 0.749 |
| ARI              | 0.624| 0.646 | 0.637 | 0.729 | 0.673 | 0.753 |
| NMI              | 0.637| 0.639 | 0.616 | 0.764 | 0.680 | 0.767 |
| **Blood**        |      |       |       |       |       |       |       |
| F1               | 0.640| 0.629 | 0.665 | 0.671 | 0.612 | 0.675 |
| Acc              | 0.643| 0.648 | 0.654 | 0.655 | 0.611 | 0.689 |
| ARI              | 0.658| 0.634 | 0.652 | 0.657 | 0.618 | 0.676 |
| NMI              | 0.649| 0.637 | 0.660 | 0.656 | 0.621 | 0.682 |
| **Ecoli**        |      |       |       |       |       |       |       |
| F1               | 0.600| 0.599 | 0.618 | 0.631 | 0.567 | 0.649 |
| Acc              | 0.565| 0.561 | 0.613 | 0.609 | 0.501 | 0.634 |
| ARI              | 0.598| 0.601 | 0.623 | 0.627 | 0.563 | 0.637 |
| NMI              | 0.612| 0.613 | 0.634 | 0.624 | 0.569 | 0.632 |
| **Banknote**     |      |       |       |       |       |       |       |
| F1               | 0.610| 0.603 | 0.620 | 0.611 | 0.631 | 0.642 |
| Acc              | 0.623| 0.599 | 0.653 | 0.620 | 0.622 | 0.650 |
| ARI              | 0.619| 0.611 | 0.616 | 0.619 | 0.624 | 0.633 |
| NMI              | 0.620| 0.607 | 0.723 | 0.615 | 0.632 | 0.737 |
| **Magic**        |      |       |       |       |       |       |       |
| F1               | 0.637| 0.629 | 0.668 | 0.679 | 0.654 | 0.777 |
TABLE 4. (Continued.) Comparison among the four benchmarks for the six clustering algorithms on the real data sets.

| Benchmark       | F1  | ARI | NMI |
|-----------------|-----|-----|-----|
|                  |     |     |     |
|                  |     |     |     |
|                  |     |     |     |
| Phoneme          | 0.674 | 0.652 | 0.729 | 0.734 | 0.698 | 0.758 |
|                  | 0.645 | 0.648 | 0.718 | 0.713 | 0.721 | 0.785 |
|                  | 0.636 | 0.643 | 0.698 | 0.708 | 0.699 | 0.773 |
|                  | 0.671 | 0.669 | 0.725 | 0.716 | 0.715 | 0.782 |
| Robotnavigation  | 0.653 | 0.652 | 0.687 | 0.673 | 0.659 | 0.698 |
|                  | 0.649 | 0.638 | 0.637 | 0.682 | 0.672 | 0.677 |
|                  | 0.620 | 0.631 | 0.645 | 0.637 | 0.653 | 0.654 |
|                  | 0.632 | 0.628 | 0.665 | 0.672 | 0.672 | 0.683 |

V. COMPLEXITY ANALYSIS

In this paper, the time complexity of the algorithm is used to describe the computational load of the algorithm, where $m$ represents the number of attributes in the numerical data set and $n$ represents the number of samples contained in the data set.

The internal coupling time complexity of the quadratic power under non-IIDness is $O(4mn)$, the time complexity of external coupling is $O(2mn)$, and the time complexity of generating the coupling representation is $O(2mn)$. The time complexity of generating the clustering results is $O(n^2)$, so the overall time complexity of this algorithm is $O(n^2)$. The time complexity of the DBSCAN algorithm is also $O(n^2)$. Although F-DBSCAN will reduce the computation time, its time complexity is still $O(n^2)$. On the test data set used in this paper, the running time of AA-DBSCAN is approximately equal to that of the DBSCAN algorithm. The time complexity of 3W-DBSCAN is mainly used to calculate the distance matrix, which also has a time complexity of $O(n^2)$. Each part of the KLS-DBSCAN algorithm has a time complexity of $O(n^2)$, so the total time complexity is $O(n^2)$. Therefore, compared with the comparison algorithms used in this paper, NDBSCAN increases the computational complexity but does not increase the time complexity of the algorithm.

Table 3 shows comparisons of the F-measure, accuracy, ARI and NMI of the 10 synthetic data sets under the proposed NDBSCAN and comparison algorithms. Fig. 5 shows the plane distributions and category labels of the 10 synthetic data sets used in this paper. Different colours represent different categories. Because this algorithm first needs to process the data and the low-dimensional data are processed into high-dimensional data, it is impossible to draw a clear cluster scatter diagram for the multidimensional data. Therefore, the clustering effects of this algorithm and the comparison algorithms are compared intuitively in Fig. 6. Table 3 and Fig. 6 show that the values of the four evaluation indexes of the proposed algorithm are higher than those of
the comparison algorithms on Smile, 2circles, Size5, and S1. The values of the F-measure, accuracy and ARI obtained by the NDBSCAN algorithm on the Moon and Flames data sets are all higher than those obtained by the comparison algorithms. The values of the F-measure, accuracy and NMI obtained by the NDBSCAN algorithm on R15 are higher than those obtained by the comparison algorithms. The advantages of the accuracy, ARI and NMI obtained by the NDBSCAN algorithm on Aggregation and Square are higher than those obtained by the comparison algorithms, indicating that the proposed algorithm can achieve good clustering effects on different types of synthetic data sets.

Table 4 shows comparisons of the four evaluation indexes of the clustering results generated by the NDBSCAN algorithm proposed in this paper and the comparison algorithms for the 11 real data sets. Fig. 7 shows the data comparison in Table 4 more intuitively. In a few cases, the evaluation values of our algorithm are slightly lower than those of the comparison algorithms. The F-measure values of the proposed algorithm on the Iris and Seeds data sets are slightly lower than those of AA-DBSCAN. The NMI value of the proposed algorithm on the Ecoli data set is slightly lower than that of 3W-DBSCAN, and the accuracy value on Banknote is slightly lower. The values of the four evaluation indexes on Sobar, Wine, Blood, Ecoli, Banknote, Magic, Phoneme and Robotnavigation of the proposed algorithm are higher than those of the comparison algorithms.

The experimental results on synthetic data sets and real data sets show that the proposed algorithm not only is suitable for small-scale low-dimensional data sets but also can achieve a better clustering effect than the comparison algorithms on large-scale low-dimensional, small-scale high-dimensional, and large-scale high-dimensional data sets. This is because the algorithm in this paper considers the internal relationship between objects and attributes of the data set and discovers the real relationship between the data by exploring the coupling relationship between the data to obtain more complete and accurate clustering results. However, this paper only focuses on numerical data, and further research is needed for categorical and mixed data.

VI. CONCLUSION
In view of the fact that there is always an interactive relationship between data and the fact that both the original and the improved DBSCAN algorithms ignore this relationship, this paper proposes a numerical DBSCAN algorithm based on the idea of a non-independent identical distribution, namely, the NDBSCAN algorithm. This algorithm calculates the relationship between data according to an improved coupling formula for calculating the relationship between numerical data, generates the coupling representation data set that can reflect the relationship between the data in the original data set, and then selects the appropriate neighbourhood parameter $\text{EPS}$ and density threshold $\text{MinPts}$ according to the distribution of the data in the coupling data set. Experimental results show that the proposed NDBSCAN algorithm is effective and superior to the comparison algorithms overall.

REFERENCES
[1] A. K. Jain, “Data clustering: 50 years beyond $K$-means,” Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, Jun. 2010.
[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” J. Roy. Stat. Soc., B, Methodol., vol. 39, no. 1, pp. 1–22, 1977.
[3] M. Girolami, “Mercer kernel-based clustering in feature space,” IEEE Trans. Neural Netw., vol. 13, no. 3, pp. 780–784, May 2002.
[4] P. Pan, H. Jiang, and Y. Lv, “Initial center optimization method of K-means algorithm within non-independent and identically distribution context,” J. Chin. Comput. Syst., vol. 40, pp. 1254–1259, Jun. 2019.
[5] L. Kaufman and P. J. Rousseeuw, “Clustering by means of medoids,” in Statistical Data Analysis Based on the L1-Norm and Related Methods, Y. Dodge, Ed. Amsterdam, The Netherlands: North Holland, 1987, pp. 405–416.
[6] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Upper Saddle River, NJ, USA: Prentice-Hall, 1988.
[7] S. C. Johnson, “Hierarchical clustering schemes,” Psychometrika, vol. 32, no. 3, pp. 241–254, Sep. 1967.
[8] S. Guha, R. Rastogi, and K. Shim, “Cure: An efficient clustering algorithm for large databases,” Inf. Syst., vol. 26, no. 1, pp. 35–58, Mar. 2001.
[9] Z. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: An efficient data clustering method for very large databases,” ACM SIGMOD Rec., vol. 25, no. 2, pp. 103–114, 1999.
[10] 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. 2nd Int. Conf. Knowl. Discovery Data Mining, Portland, OR, USA, Aug. 1996, pp. 226–231.
[11] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, “OPTICS: Ordering points to identify the clustering structure,” ACM SIGMOD Rec., vol. 28, no. 2, pp. 49–60, Jun. 1999.
[12] A. Hinneburg and D. A. Keim, “An efficient approach to clustering in large multimedia databases with noise,” in Proc. 4th Int. Conf. Knowl. Discovery Data Mining, vol. 98, Aug. 1998, pp. 58–65.
[13] E. Schikuta, “Grid-clustering: An efficient hierarchical clustering method for very large data sets,” in Proc. 13th Int. Conf. Pattern Recognit., vol. 2, Aug. 1996, pp. 101–105.
[14] E. Schikuta and M. Erhart, “The BANG-clustering system: Grid-based data analysis,” in Proc. Int. Symp. Intell. Data Anal., vol. 1280, Berlin, Germany: Springer, Aug. 1997, pp. 513–524.
[15] W. Wang, J. Yang, and R. Munzt, “STING: A statistical information grid approach to spatial data mining,” in Proc. 23rd Int. Conf. Very Large Data Bases (VLDB), Athens, Greece, Aug. 1997, pp. 1–18.
[16] M. Pavlisa, L. Dolega, and A. Singleton, “A modified DBSCAN clustering method to estimate retail center extent,” Geograph. Anal., vol. 50, no. 2, pp. 141–161, Sep. 2017.
[17] J. Zhu, H. Zeng, J. Huang, S. Liao, Z. Lei, C. Cai, and L. Zheng, “Vehicle re-identification using quadruple directional deep learning features,” IEEE Trans. Intell. Transp. Syst., vol. 21, no. 1, pp. 410–420, Jan. 2020.
[18] S. Pei, T. Shen, X. Wang, C. Gu, Z. Ning, X. Ye, and N. Xiong, “3DACN: 3D augmented convolutional network for time series data,” Inf. Sci., vol. 513, pp. 17–29, Mar. 2020.
[19] S. Liu, J. Wang, Y. Lu, H. Li, J. Zhao, and Z. Zhu, “Multi-focus image fusion based on adaptive dual-channel spiking cortical model in non-subsampled shearlet domain,” IEEE Access, vol. 7, pp. 56367–56388, 2019.
[20] H. Chebi, D. Acheli, and M. Kesraoui, “Dynamic detection of abnormalities in video analysis of crowd behavior with DBSCAN and neural networks,” Adv. Sci., Technol. Eng. Syst. J., vol. 1, no. 5, pp. 56–63, Oct. 2016.
[21] K. Sheridan, “An application of DBSCAN clustering for flight anomaly detection during the approach phase,” in Proc. AIAA Scitech Forum, Jan. 2020, p. 1851.
[22] Z.-H. Feng, X.-Z. Qian, and N.-N. Zhao, “Greedy DBSCAN: An improved DBSCAN algorithm on multi-density clustering,” Appl. Res. Comput., vol. 33, no. 9, pp. 2693–2696, 2016.
[23] M. M. R. Khan, M. A. B. Siddique, R. B. Arif, and M. R. Oishe, “ADBSCAN: Adaptive density-based spatial clustering of applications with noise for identifying clusters with varying densities,” in Proc. 4th Int. Conf. Electr. Eng. Inf. Commun. Technol. (iCEEiCT), Sep. 2018, pp. 107–111.
[24] A. Starczewski, P. Goetzen, and M. J. Er, “A new method for automatic determining of the DBSCAN parameters,” J. Artif. Intell. Soft Comput. Res., vol. 10, no. 3, pp. 209–221, Jul. 2020.

[25] P. Lin, Z. Hong, W. Feng, Y. Li, and L. Wu, “Design and implementation of an improved DBSCAN algorithm,” in Proc. IEEE 3rd Adv. Inf. Manage., Commun., Electron. Automat. Control Conf. (IMCEC), Oct. 2019, pp. 1834–1839.

[26] H. Yu, L. Chen, J. Yao, and X. Wang, “A three-way clustering method based on an improved DBSCAN algorithm,” Phys. A, Stat. Mech. Appl., vol. 555, Dec. 2019, Art. no. 122289.

[27] J.-H. Kim, J.-H. Choi, K.-H. Yoo, and A. Nasridinov, “AA-DBSCAN: An approximate adaptive DBSCAN for finding clusters with varying densities,” J. Supercomput., vol. 75, no. 1, pp. 142–169, Jan. 2019.

[28] W. Guang and L. Guoyu, “Improved adaptive parameter DBSCAN clustering algorithm,” Comput. Eng. Appl., vol. 56, pp. 45–51, 2020.

[29] L. Yikun, “Research on DBSCAN algorithm under non-independent and identical distribution,” M.S. thesis, Dept. Electron. Eng., Qilu Univ. Technol., Jinan, China, 2020.

[30] L. Cao, “Non-IIDness learning in behavioral and social data,” Comput. J., vol. 57, no. 9, pp. 1358–1370, Sep. 2014.

[31] YANMEI WANG received the bachelor’s degree from the Qilu University of Technology (Shandong Academy of Sciences), in 2019, where she is currently pursuing the master’s degree. Her research interests include data mining and big data processing.

HE JIANG received the B.Sc. degree from Liaocheng University, in 1985, and the M.E. degree in computer applications from Beijing Jiaotong University, in 1997. He was an Associate Professor with the Qilu University of Technology, in 2001, and a Professor with the Qilu University of Technology, in 2005. He served as the Director of the School of Computer Science and Technol- ogy, Qilu University of Technology, for a long time. He was also the Director of the Computer Teaching Professional Research Institute, Shandong Higher Education Association. At Qilu University of Technology, he was the first Famous Teacher and a Teacher Morality Model. His research interests include data mining technology, big data analysis, and other aspects of research.

CHENGQIANG WANG received the bachelor’s degree from the Qilu University of Technology (Shandong Academy of Sciences), in 2019. His research interests include data mining and data security.

[32] C. Wang, Z. She, and L. Cao, “Coupled attribute analysis on numerical data,” in Proc. Int. Joint Conf. Artif. Intell. (IJCAI), Aug. 2013, pp. 1736–1742.

[33] G. Gan, C. Ma, and J. Wu, Data Clustering: Theory, Algorithms, and Applications (ASA-SIAM Series on Statistics and Applied Probability). 2013.

[34] (2016). Machine Learning Clustering Datasets. [Online]. Available: http://cs.joensuu.fi/sipu/datasets

[35] Build Software Better, Together. [Online]. Available: https://github.com

[36] K. Bache and M. Lichman, “UCI machine learning repository,” School Inf. Comput. Sci., Univ. California, Irvine, Irvine, CA, USA, Tech. Rep., 2013.

[37] D. Cai, X. He, and J. Han, “Document clustering using locality preserving indexing,” IEEE Trans. Knowl. Data Eng., vol. 17, no. 12, pp. 1624–1637, Dec. 2005.

[38] J. Lu and Q. Zhu, “An effective algorithm based on density clustering framework,” IEEE Access, vol. 5, pp. 4991–5000, 2017.

[39] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada, “Normalized mutual information feature selection,” IEEE Trans. Neural Netw., vol. 20, no. 2, pp. 189–201, Feb. 2009.

[40] A. Lancichinetti, S. Fortunato, and J. Kertész, “Detecting the overlapping and hierarchical community structure in complex networks,” New J. Phys., vol. 11, no. 3, Mar. 2009, Art. no. 033015.