Research Article

A Clustering Algorithm via Density Perception and Hierarchical Aggregation Based on Urban Multimodal Big Data for Identifying and Analyzing Categories of Poverty-Stricken Households in China

Hui Liu ^{1,2}, Yang Liu ³, Ran Zhang ¹, and Xia Wu ⁴

¹Kaifu Campus of Dalian University of Technology, No. 321, Tuqiang Street, Dalian Economic and Technological Development Zone, Dalian, Liaoning 116600, China

²Faculty of Business and Management, Universiti Teknologi MARA, Cawangan Sarawak, Jalan Meranek, 94300 Kota Samarahan, Sarawak, Malaysia

³International School of Shenyang Jianzhu University, No. 25, Hunnan East Road, Hunnan New District, Shenyang, Liaoning 110168, China

⁴Huawei Nanjing Research & Development Center, No. 101 Software Avenue, Yuhuatai District, Nanjing, Jiangsu 210012, China

Correspondence should be addressed to Yang Liu; liuyang@sjzu.edu.cn

Received 11 November 2020; Revised 28 December 2020; Accepted 2 February 2021; Published 23 February 2021

Academic Editor: Liang Zou

Copyright © 2021 Hui Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Nowadays, urban multimodal big data are freely available to the public due to the growing number of cities, which plays a critical role in many fields such as transportation, education, medical treatment, and land resource management. The successful completion of poverty-relief work can greatly improve the quality of people's life and ensure the sustainable development of the society. Poverty is a severe challenge for human society. It is of great significance to apply machine learning to mine different categories of poverty-stricken households and further provide decision support for poverty alleviation. Traditional poverty alleviation methods need to consume a lot of manpower, material resources, and financial resources. Based on the density-based spatial clustering of applications with noise (DBSCAN), this paper designs the hierarchical DBSCAN clustering algorithm to identify and analyze the categories of poverty-stricken households in China. First, the proposed method adjusts the neighborhood radius dynamically for dividing the data space into several initial clusters with different densities. Then, neighbor clusters are identified by the border and inner distances constantly and aggregated recursively to form new clusters. Based on the idea of division and aggregation, the proposed method can recognize clusters of different forms and deal with noises effectively in the data space with imbalanced density distribution. The experiments indicate that the method has the ideal performance of clustering, which identifies the commonness and difference in characteristics of poverty-stricken households reasonably. In terms of the specific indicator "Accuracy," the accuracy increases by 2.3% compared with other methods.

1. Introduction

With the development of Information and Communication Technology, the era of multimodal big data has arrived comprehensively. Cities are the important places which are of prime importance for big data distribution, such as population, economy, transportation, and landscape [1–3]. The urban multimodal big data obtained by traditional data

collection methods such as field survey and questionnaire interview cannot objectively and accurately reflect the status quo of urban development and the law of residents' activities in a wide range of time and space. Also, the obtained urban operation information has a large lag. Multimodal big data can make up for the above defects and deeply depict the urban physical space and social environment. This not only provides the possibility to objectively understand the urban

system and summarize its development rules but also provides important support for urban planning and related research studies such as poverty-relief work and urban education.

It must be admitted that urban planning based on urban multimodal big data is a very challenging task for poverty-relief work. It can improve urban environments, quality of life, and smart city systems [4, 5]. Due to the short time, heavy task of targeted poverty alleviation in the early stage, the basic information of each impoverished object and the causes of poverty are not comprehensive and accurate enough, which needs to be further enriched and improved. Poor object management mechanism is not perfect. Due to the large number of poor people in the poor villages and the complicated family situation, the number of people coming out of the basin and returning to poverty is in constant change [6]. In addition, the management mechanism of poor objects at the village level is not sound enough, so there is a lack of changes in the poor population in the poor villages.

In this paper, we focus on the tasks of identifying and analyzing categories of poverty-stricken households in China. Eradication of poverty is the historical task facing the international community. With the development of artificial intelligence (AI) technologies such as machine learning and deep learning, a growing number of researchers are making great efforts to develop and unleash the huge potential of these AI technologies in alleviating poverty [7]. China, as the largest developing country worldwide, has made a significant contribution to global poverty alleviation. In the year of 2013, the Chinese government raised the concept of targeted poverty alleviation, which aims to take targeted measures to assist each truly poverty-stricken household and eliminate various factors leading to poverty fundamentally, thus achieving the goal of sustainable poverty alleviation [8]. On the basis of the policy, this paper adopts the clustering algorithm [9] to divide the data of poverty-stricken households in China reasonably and thus identify different categories of poverty-stricken households for supporting the formulation and implementation of antipoverty measures.

Poverty-oriented scientific research depends on the analysis of poverty data. The Chinese poverty data generally come from population censuses carried out by the country, society, and universities [10]. Due to the wide coverage of population and the individual differences in educational level and psychology, respondents may not answer questionnaires according to actual conditions, which results in the subjectivity of questionnaire data. Additionally, faults in processes such as data entry and storage can easily lead to outliers and missing values in datasets. Since the quality of poverty datasets obtained by population censuses is hard to guarantee, it brings certain difficulties for the design and application of clustering algorithms.

The design of clustering algorithms for poverty datasets should make reasonable consideration of noises caused by missing values and outliers. Nowadays, common clustering methods mainly include partitional clustering, hierarchical clustering, and density-based clustering [11]. The K-means clustering algorithm achieves clustering through the partition, which assigns each sample to the closest cluster

according to distances between samples and prototypes and updates prototypes by the average of samples within clusters, then repeats the above steps until the iteration ends [12]. Although the method is easy and practicable, the number of clusters and the initial prototypes need to be predefined. The agglomerative hierarchical clustering (AHC) regards each sample as a separate cluster and then merges the two closest clusters into a new cluster constantly [13]. The AHC algorithm requires no predefined prototypes and can get the hierarchical structure of clusters, but it is sensitive to noises within data. The density-based spatial clustering of applications with noise (DBSCAN) algorithm is a representative of density-based clustering methods, which defines the cluster as the maximal set of density-connected samples and takes the sample regions with high densities as clusters, thus discovering clusters of arbitrary shapes [14] whereas the hyperparameters ϵ and minpts in the DBSCAN algorithm, i.e., the neighborhood radius and the minimum number of samples required to form a dense region, have a great influence on the result of clustering, and the method is not applicable to datasets with different density distribution. Many researchers improve DBSCAN in view of the existing problems in the algorithm and propose improved algorithms such as K-nearest neighbor DBSCAN (KNNDSCAN), DVSCAN, and varied density-based spatial clustering of applications with noise (VDBSCAN) [15–18]. For instance, Gaonkar and Sawant [19] drew a k-dist graph based on the distance between each sample and its k-th nearest neighbor so as to identify multiple values of the neighborhood radius, then finds the clusters with different densities under each value of the neighborhood radius. Fahim et al. proposed an enhanced DBSCAN (EDSCAN) algorithm, which defined the density variation for core points and specified that a core point allowed for expansion only when its density variation was less than or equal to a threshold value and its neighborhood satisfies the homogeneity index [20]. In terms of the clustering methods, some other researchers proposed many advanced approaches such as robust FCM clustering [21], improved quantum clustering algorithm [22], and swarm clustering algorithm [23]. Chen et al. [24] proposed a fast clustering for large-scale data. Chel et al. [25] presented the HDBSCAN clustering algorithm to find a clustering pattern present in calcium spiking obtained by confocal imaging of single cells. Znidi et al. [26] introduced a new methodology for discovering the degree of coherency among buses using the correlation index of the voltage angle between each pair of buses and used the hierarchical density-based spatial clustering of applications with noise to partition the network into islands. Parmar et al. [27] proposed a residual error-based density peak clustering algorithm named REDPC to better handle datasets comprising various data distribution patterns. Specifically, REDPC adopted the residual error computation to measure the local density within a neighborhood region. Parmar et al. [28, 29] proposed the feasible residual error-based density peak clustering algorithm with the fragment merging strategy, where the local density within the neighborhood region was measured through the residual error computation and the resulting residual errors were then used to generate residual fragments for cluster formation.

Overall, the above methods have the limits of low clustering efficiency and time-consuming with high-dimensional data.

Considering that clusters in real-world datasets may have different sizes, shapes, and densities, accompanied by certain noises and outliers, this paper takes the idea of initial division and hierarchical aggregation to design a clustering algorithm named hierarchical DBSCAN (HDBSCAN). The proposed method comprises two stages of division and aggregation. Our contributions are as follows:

- (1) First, it makes an initial division of the dataset based on sample densities; that is, the proposed method takes the neighbor information of samples to calculate local density values and then searches the set of density-connected samples for each unlabeled core point sequentially according to the density values in descending order, thus forming the initial clusters.
- (2) Then, the method adopts the idea of hierarchical clustering to perform the aggregation of neighbor clusters. Based on the inner and border distances between clusters, the most similar clusters are regarded as neighbor clusters and merged to form a new cluster, and the process is repeated until the iteration ends.
- (3) Based on the way of division and aggregation, the method can identify clusters with different forms in the dataset. Moreover, noise data cannot be integrated into high-density clusters as its density is relatively sparse, by which the proposed method can handle noise data reasonably.

The rest of this paper is organized as follows. Section 2 introduces two typical clustering algorithms, i.e., the DBSCAN clustering and the hierarchical clustering. Section 3 describes the proposed hierarchical DBSCAN algorithm in detail. Section 4 discusses the clustering performance of the proposed method, then applies it to the Chinese poverty dataset, and further analyzes the result of clustering. Finally, conclusions are presented in Section 5.

2. Theoretical Foundation

2.1. The DBSCAN Clustering. The DBSCAN algorithm regards regions with high densities as clusters and those with sparse densities as noises. It requires two hyperparameters, i.e., the neighborhood radius eps and the minimum number of samples required to form a dense region $minpts$.

Let $D = \{x_1, \dots, x_n\}$ represent the dataset composed of n samples and d attributes, where $x_i = [x_{i1}, \dots, x_{id}]^T$ denotes the i -th sample in the dataset. The eps -neighborhood of x_i is

$$N_{eps}(x_i) = \{x_j \in D | \text{dist}(x_i, x_j) \leq eps\}, \quad (1)$$

where $\text{dist}(x_i, x_j)$ denotes the distance between samples x_i and x_j , calculated by

$$\text{dist}(x_i, x_j) = \sqrt{\sum_{k=1}^d (x_{ik} - x_{jk})^2}. \quad (2)$$

If x_i satisfies equation (3), it is called the core point:

$$|N_{eps}(x_i)| \geq minpts. \quad (3)$$

There are several definitions in the DBSCAN algorithm, listed as follows:

- (1) A sample x_j is directly reachable from x_i with respect to eps and $minpts$ if x_i is a core sample and $x_j \in N_{eps}(x_i)$
- (2) A sample x_j is reachable from x_i with respect to eps and $minpts$ if there exists a chain of samples $x_{m_1}, \dots, x_{m_k}$, ($1 \leq k, m_k \leq n$) with $x_{m_1} = x_i$ and $x_{m_k} = x_j$, where each $x_{m_{l+1}}$ ($0 < l < k$) is directly reachable from x_{m_l} with respect to eps and $minpts$
- (3) A sample x_j is reachable from x_i with respect to eps and $minpts$ if there exists a chain of samples $x_{m_1}, \dots, x_{m_k}$, ($1 \leq k, m_k \leq n$) with $x_{m_1} = x_i$ and $x_{m_k} = x_j$, where each $x_{m_{l+1}}$ ($0 < l < k$) is directly reachable from x_{m_l} with respect to eps and $minpts$

In the process of clustering, the algorithm randomly selects a core point as the initial point and takes all the core points in its eps -neighborhood for continuous expansion. The expansion ends until the maximal set of density-connected samples is found and labeled as one cluster. After that, the algorithm randomly chooses other unlabeled core points for generating new clusters. The process of clustering completes when all the core points are labeled.

2.2. Hierarchical Clustering. The hierarchical clustering can be divided into the agglomerative hierarchical clustering and the divisive hierarchical clustering. The agglomerative hierarchical clustering first takes each sample as a separate cluster, then finds the two closest clusters by measuring the distance between the clusters, and then merges them into a new cluster. Subsequently, the algorithm recalculates the distance between clusters and continues the aggregation process. The realization of the divisive hierarchical clustering is the exact opposite of the above, which regards the whole dataset as one cluster and then performs the division iteratively.

In the hierarchical clustering, the distance between C_p and C_q can be calculated by (4), i.e., the average of sample distances between two clusters. Besides, the minimum distance of samples between clusters shown in (5), or the maximum distance of samples between clusters, can also be taken to measure the distance of two clusters:

$$S_1(C_p, C_q) = \frac{1}{|C_p| \cdot |C_q|} \sum_{x_i \in C_p} \sum_{x_j \in C_q} \text{dist}(x_i, x_j), \quad (4)$$

$$S_2(C_p, C_q) = \min\{\text{dist}(x_i, x_j) | x_i \in C_p, x_j \in C_q\}. \quad (5)$$

2.3. The Hierarchical DBSCAN Algorithm. As the global hyperparameters for the DBSCAN algorithm, the numerical values of minpts and eps have a direct impact on the expansion of all the clusters. Figure 1 illustrates the expansion of clusters under different numerical values of eps, where the red points denote the initial core points in each iteration of expansion. According to Figure 1(a), the clusters C_1 and C_2 can be identified while the other samples are regarded as noises and cannot be partitioned properly if the DBSCAN algorithm takes eps_1 as the neighborhood radius. It can be seen from Figure 1(b) that all the samples are divided into one cluster C_1 through four iterations of expansion if the algorithm takes eps_2 as the neighborhood radius.

In view of the above problem, this paper takes the way of division and aggregation to design the HDBSCAN clustering algorithm. First, the proposed method makes an initial division of the dataset according to sample densities. During the expansion of each cluster, the method adaptively adjusts the neighborhood radius based on the neighbor information of samples within the cluster. Then, the idea of hierarchical clustering is adopted to perform the recursive aggregation; that is, the method takes the cluster pair with the minimum distance as the neighbor clusters and then merges them into a new cluster. Based on division and aggregation, the method can perceive the clusters with different forms in the data space.

2.4. Initial Division. During the process of initial division, the parameter k is used to calculate the local density. Let $\text{SN}_k(x_i)$ represent the set composed of k samples closest to x_i , and the average distance between x_i and all samples in the set is

$$\overline{\text{dist}}(x_i) = \frac{1}{|\text{SN}_k(x_i)|} \sum_{x_j \in \text{SN}_k(x_i)} \text{dist}(x_i, x_j). \quad (6)$$

The distance $\overline{\text{dist}}(x_i)$ can capture the density distribution around the sample x_i . The smaller the value, the greater the density. Therefore, the local density of x_i can be defined as

$$\text{den}(x_i) = \overline{\text{dist}}(x_i)^{-1}. \quad (7)$$

The neighborhood radius of x_i , namely, $\text{eps}(x_i)$, is the distance between x_i and the maxpts-th nearest sample. The process of the initial division includes the following steps.

Step 1. Calculate the local density for each sample and then sort the samples based on the local density values so as to form the sequence:

$$O = \left\{ x_{m_1}, \dots, x_{m_n} \mid \text{den}(x_{m_j}) > \text{den}(x_{m_{j+1}}), 1 \leq j \leq n-1, 1 \leq m_j \leq n \right\}. \quad (8)$$

The cluster label is initialed as $q = 1$.

Step 2. Select an unlabeled sample x_i from the sequence O in order and set the iteration number $t = 1$.

Step 3. Let $C_q^{(t)}$ and $Q_q^{(t)}$ represent the set of samples and the sequence of core points for the q -th cluster in the t -th iteration and $C_q^{(1)} = \{x_i\}$, $Q_q^{(1)} = \{x_i\}$.

Step 4. Calculate the adaptive neighborhood radius for the expansion of the current cluster by all samples in the cluster:

$$\text{eps}(C_q^{(t)}) = \frac{1}{|C_q^{(t)}|} \sum_{x_k \in C_q^{(t)}} \text{eps}(x_k). \quad (9)$$

Step 5. Select a core point x_j from the sequence $Q_q^{(t)}$ in order and continue the expansion based on $\text{eps}(C_q^{(t)})$.

Step 6. Calculate the set of neighbor samples to be expanded according to

$$\begin{aligned} \widehat{C}_q^{(t)} &= \{x_k \mid \text{dist}(x_j, x_k) \\ &\leq \text{eps}(C_q^{(t)}) \text{ and } x_k \notin C_p, p = 1, \dots, q-1\}. \end{aligned} \quad (10)$$

Step 7. Update $C_q^{(t+1)}$ and $Q_q^{(t+1)}$ by

$$C_q^{(t+1)} = C_q^{(t)} \cup \widehat{C}_q^{(t)}, \quad (11)$$

$$\begin{aligned} Q_q^{(t+1)} &= Q_q^{(t)} \cup \left\{ x_k \mid x_k \in \widehat{C}_q^{(t)} \mid N_{\text{eps}(C_q^{(t)})}(x_k) \right. \\ &\quad \left. \geq \text{minpts} \right\} - \{x_j\}. \end{aligned} \quad (12)$$

Step 8. The expansion of the q -th cluster C_q is completed if $Q_q^{(t+1)} = \emptyset$, then it returns to Step 9. Otherwise, it sets $t = t + 1$ and returns to Step 4.

Step 9. The initial division ends if all the samples are labeled. Otherwise, it sets the cluster label as $q = q + 1$ and returns to Step 2.

2.5. Aggregation of Neighbor Clusters. In this paper, the similarity between clusters is measured by border distance and inner distance. Figure 2 takes the clusters C_p' and C_q' during the aggregation as an example to describe two kinds of distances. In Figure 2, the red points denote the core points and the grey ones denote the border points distributed around the clusters.

Suppose that the dataset can be represented by $D = \{C_1, \dots, C_K\}$ after the initial division, where K denotes the number of clusters and C_i ($i = 1, \dots, K$). While the neighbor clusters are merged to form new clusters continuously during the aggregation, C_p' is described by $C_p' = \{C_{m_1}, \dots, C_{m_{K_p}}\}$, $1 \leq K_p, m_{K_p} \leq K$. The set of border points in C_p' is

$$B(C_p') = \left\{ x_i \mid x_i \in C_p' \text{ and } |N_{\text{eps}(x_i)}(x_i)| < \text{minpts} \right\}, \quad (13)$$

where $\text{eps}(x_i)$ denotes the neighborhood radius at the completion of division for x_i . The value changes dynamically due

FIGURE 1: The expansion of clusters under different numerical values of eps: (a) expansion with eps_1 ; (b) expansion with eps_1 .

FIGURE 2: The border distance and the inner distance: (a) border distance; (b) inner distance.

to the adaptive adjustment of neighbor radius. According to Figure 2(a), the border distance between clusters C'_p and C'_q is the minimum distance between the border points of two clusters, namely,

$$O(C'_p, C'_q) = \min\{\text{dist}(x_i, x_j) | x_i \in B(C'_p), x_j \in B(C'_q)\}. \quad (14)$$

As can be seen from Figure 2(b), the cluster C'_p consists of four initial clusters, and thus the inner distance of the cluster is defined as

$$I(C'_p) = \frac{2}{K_p(K_p - 1)} \sum_{C_{m_i} \in C'_p} \sum_{C_{m_j} \in C'_p, C_{m_j} \neq C_{m_i}} O(C_{m_i}, C_{m_j}). \quad (15)$$

During the aggregation, the two clusters with the minimum border distance are considered as the neighbor clusters for further merging if their difference of inner distances and that of densities below certain limitations. Algorithm 1 is a simple implementation of aggregation for neighbor clusters. In the actual implementation of the algorithm, values such as border distances and inner distances will be restored to avoid

repeated calculation. According to the 14th line of Algorithm 1, two clusters will be involved in calculating neighbor clusters only when their density difference, border distances, and inner distances satisfy certain conditions.

The proposed HDBSCAN clustering algorithm can capture clusters with different forms in the data space. The aggregation of neighbor clusters weakens the sensitivity of the algorithm to hyperparameters in the initial division. Besides, the result of the division in the DBSCAN algorithm depends on the selection sequence of initial core points. The proposed method can weaken the fluctuation caused by the selection sequence to some extent. The Algorithm 2 summarizes the whole process.

3. Experimental Results and Analysis

3.1. Experimental Design

3.1.1. Datasets. Three public artificial datasets and four real-world datasets are chosen to verify the effectiveness of the proposed clustering algorithm. The description of artificial datasets is listed in Table 1. The visualization of artificial datasets is shown in Figure 3.

```

(1) Input: clusters after initial division  $D = \{C_1, \dots, C_K\}$ ; the threshold  $\varepsilon$ ;  $\{\text{den}(x_i) | i = 1, \dots, n\}$ 
(2) Output: final clusters after aggregation  $D = \{C'_1, \dots, C'_K\}$ 
(3)  $\text{min\_O} \leftarrow +\infty, \text{combine} \leftarrow \emptyset$ 
(4) While True
(5)   Calculate  $\text{average\_den\_diff}$  by the averaged of density differences between clusters
(6)   For each cluster  $C'_p$  in  $D$ 
(7)     For each cluster  $C'_q$  in  $D - C'_p$ 
(8)       Calculate  $O(C'_p, C'_q), I(C'_p), I(C'_q)$ 
(9)       Calculate  $\text{den}(C'_p), \text{den}(C'_q)$  by the averaged densities for samples in the clusters
(10)     $\text{den\_diff} \leftarrow |\text{den}(C'_p) - \text{den}(C'_q)| / \max(\text{den}(C'_p), \text{den}(C'_q))$ 
(11)     $\text{tmp} = \max(O(C'_p, C'_q), I(C'_p), I(C'_q))$ 
(12)     $\text{dist\_diff} \leftarrow ((O(C'_p, C'_q) - \max(I(C'_p), I(C'_q))) / \text{tmp})$ 
(13)     $O \leftarrow O(C'_p, C'_q)$ 
(14)    If  $\text{den\_diff} < \text{average\_den\_diff}$  and  $\text{dist\_diff} < \varepsilon$  and  $O < \text{min\_O}$ 
(15)       $\text{min\_O} \leftarrow O, \text{combine} \leftarrow \{C'_p, C'_q\}$ 
(16)    End For
(17)  End For
(18)  If  $\text{combine} \neq \emptyset$ 
(19)     $D \leftarrow D - \text{combine}[0] - \text{combine}[1] + \{\text{combine}[0] \cup \text{combine}[1]\}$ 
(20)    Else
(21)      Break
(22) End While

```

ALGORITHM 1: The aggregation of neighbor clusters.

```

(1) Input: parameter  $k$ , clusters after initial division  $D = \{C_1, \dots, C_K\}$ ; the threshold  $\varepsilon$ ;  $\{\text{den}(x_i) | i = 1, \dots, n\}$ 
(2) Output: final clusters
(3)  $\text{min\_O} \leftarrow +\infty, \text{combine} \leftarrow \emptyset$ 
(4) While True
(5)   Calculate the local density
        $O = \{x_{m_1}, \dots, x_{m_n} | \text{den}(x_{m_j}) > \text{den}(x_{m_{j+1}}), 1 \leq j \leq n-1, 1 \leq m_j \leq n\}$ 
(6)   Select an unlabeled sample  $x_i$  from the sequence  $O$ 
(7)   For  $C_q^{(1)} = \{x_i\}, Q_q^{(1)} = \{x_i\}$ .
(8)     Calculate the adaptive neighborhood radius
(9)     Select a core point  $x_j$  from the sequence  $Q_q^{(t)}$ 
(10)    Calculate the set of neighbor samples
(11)    For  $Q_q^{(t+1)} = \emptyset$ 
(12)      The expansion of the  $q$ -th cluster  $C_q$  is completed
(13)    End For
(14)  End For
(15)  Calculate  $\text{average\_den\_diff}$  by the averaged of density differences between clusters
(16)  If  $\text{den\_diff} < \text{average\_den\_diff}$  and  $\text{dist\_diff} < \varepsilon$  and  $O < \text{min\_O}$ 
(17)     $\text{min\_O} \leftarrow O, \text{combine} \leftarrow \{C'_p, C'_q\}$ 
(18)  End For
(19)  End For
(20)  If  $\text{combine} \neq \emptyset$ 
(21)     $D \leftarrow D - \text{combine}[0] - \text{combine}[1] + \{\text{combine}[0] \cup \text{combine}[1]\}$ 
(22)  Else
(23)    Break
(24) End While

```

ALGORITHM 2: The proposed cluster method.

The description of real-world datasets is listed in Table 2, where Banknote, Parkinson, Codon usage, HCV, and Planning relax are taken from UCI machine learning repository, and CFPS2016 is the dataset of poverty-stricken households in China. The CFPS2016 dataset comes from the

China Family Panel Studies (CFPSs) released by the Institute of Social Science Survey of Peking University, China, in 2016. In the experiment, the CFPS2016 dataset consists of 14019 samples and 320 attributes, which covers the family economy as well as the states of adults and children in health,

TABLE 1: The description of artificial datasets.

Data sets	Size	Dimension	Cluster number
cluto-t5-8k	8000	2	6
cluto-t8-8k	8000	2	8
triangle2	8000	2	10

FIGURE 3: The visualization of three artificial datasets: (a) cluto-t5-8k, (b) cluto-t8-8k, and (c) triangle2.

TABLE 2: The description of real-world datasets.

Data sets	Records	Attributes
Banknote	1372	5
Parkinson	1040	26
Planning relax	182	13
Codon usage	13028	69
HCV	615	14
CFPS2016	1778	320

education, and psychology. Hence, the CFPS2016 dataset can reflect the status of each Chinese household objectively. During the data preprocessing, we fill in missing values with the K-nearest neighbor imputation method [30], and then 1778 poverty-stricken households are measured from 14019 Chinese households based on the Alkire–Foster method, the main measurement method of multidimensional poverty [31]. The parameters in this experiment are set the same as DBSCAN under the same experimental platform.

3.1.2. Evaluation Metrics. We take the silhouette coefficient (SC) [32], Davies–Bouldin index (DBI) [33], adjusted Rand index (ARI), and normalized mutual information (NMI) [34] to measure the performance of clustering. The silhouette coefficient is defined by

$$SC = \frac{1}{n} \sum_{i=1}^n \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (16)$$

where n denotes the total number of samples; $a(i)$ denotes the average distance between the sample x_i and all other samples in its cluster, which reflects the cohesiveness of clustering; and $b(i)$ denotes the minimum value of average distances between the sample x_i and all samples in any other cluster, which reflects the dispersity of clustering. The larger SC represents the higher performance of clustering. Besides, the definition of the Davies–Bouldin index is

$$DBI = \frac{1}{K'} \sum_{i,j=1}^{K'} \max_{i \neq j} \left(\frac{\bar{S}_i - \bar{S}_j}{\|w_i - w_j\|_2} \right), \quad (17)$$

where K' denotes the number of clusters; $\bar{S}_i$ and $\bar{S}_j$ denote the average distance between all the samples within the cluster and the centroid of the cluster; $\|w_i - w_j\|_2$ denotes the distance between cluster centroids. The smaller DBI denotes the higher performance of clustering.

With respect to performance, adjusted Rand index (ARI) and normalized mutual information (NMI) are also used for evaluation. ARI represents the similarity measure between two clusterings that is adjusted for chance and is related to accuracy, while NMI quantifies the amount of information obtained about one clustering, through the other clustering (i.e., the mutual dependence between the two). In the case of observations being identified as noise, each noise observation is treated as a distinct singleton cluster for both ARI and NMI.

3.1.3. Compared Methods. This paper compares the proposed method with three existing clustering algorithms which are described as follows:

- (1) AHC: as described in Section 2.2, the method regards every sample as a separate cluster and then merges the two closest clusters continuously until the iteration ends.
- (2) DBSCAN: as described in Section 2.1, the method performs the continuous expansion for each cluster based on core points and thus takes regions with high densities as clusters and those with low-densities as noises.
- (3) EDBSCAN: the method calculates the density variation for each core points and specified that a core point is allowed to expand only when its density variation is below a specified threshold and its neighborhood satisfies the homogeneity index [35].
- (4) NS-DBSCAN: the NS-DBSCAN algorithm used a strategy similar to the DBSCAN algorithm. Furthermore, it provided a new technique for visualizing the density distribution and indicating the intrinsic clustering structure [36].
- (5) ADBSCAN: unlike many other algorithms that estimate the density of each samples using different kinds of density estimators and then choose core samples based on a threshold, ADBSCAN utilized the inherent properties of the nearest neighbor graph [37].

4. Results and Analysis

4.1. Artificial Datasets and Real-World Datasets from UCI. First, we conduct the effect experiments of ε on the local sensitivity as shown in Figure 4. Then, the selected ε is used for the following experiments to provide the equitable comparison. From Figure 4, we can know that when ε is 0.5, the local sensitivity is small. The effect of proposed method is better. Therefore, we select $\varepsilon = 0.5$ in this paper.

The clustering results of three artificial datasets based on the proposed method are shown in Figure 5, where regions

with different colors can be regarded as one cluster. According to Figures 5(a), 5(c), and 5(e), the datasets are cut into several regions with different densities after the initial division. As can be seen from Figures 5(b), 5(d), and 5(f), the adjacent regions with similar densities aggregate continuously during the aggregation of neighbor clusters, which contributes to the ideal results of clustering. In Figure 5(f), some discrete points are distributed around four large clusters. The proposed method identifies these points as noises since there exist certain differences between the densities of discrete points and those of clusters around them.

The metric values for three UCI datasets obtained by four comparison methods are shown in Table 3, in which the optimal results have been bolded and the suboptimal results have been italicized.

According to Table 2, all the SC values obtained by the proposed method HDBSCAN are better than those obtained by other methods, and the method also has ideal DBI values. For instance, in respect of the Parkinson dataset, the SC value of HDBSCAN is 8.91% higher than that of the suboptimal method AHC. Although the DBI value of HDBSCAN is suboptimal, it is only 2.63% worse than that of EDBSCAN. The above results indicate that the proposed method HDBSCAN has the ideal performance of clustering. Table 2 shows the ARI performance with the different methods on the artificial datasets. From these results, HDBSCAN is shown to rank first in these datasets. More importantly, in each case HDBSCAN is able to identify the underlying classes of each dataset, whereas each of the other approaches fails at this task in at least one case.

4.2. The Dataset of Poverty-Stricken Households in China. We perform clustering on 1778 poverty-stricken households of CFPS2016 so as to identify different categories of poverty-stricken households. Table 4 shows the metric values for CFPS2016 obtained by four compared methods, where the optimal results have been bolded and the suboptimal results have been italicized. Table 4 also shows NMI performance results on the same set of artificial datasets and clustering approaches. Here, HDBSCAN ranking performance is identical to those discussed with respect to ARI.

We also make accuracy comparison with the other three methods. The results are the average values shown in Table 5.

It can be seen from Table 5 that the values of SC and DBI obtained by HDBSCAN are better than those obtained by other compared methods. Therefore, the proposed method has the ideal performance of clustering on the CFPS2016 dataset. The clustering result based on HDBSCAN is listed in Table 6.

According to Table 6, the proposed method divides CFPS2016 into 10 clusters and identifies 70 noises. Additionally, the numbers of households within different clusters are distributed unevenly. For instance, the number of households in Cluster 1 is 382 while those in Cluster 9 and Cluster 10 are 61 and 34, respectively. To evaluate the rationality of the clustering result, we adopt the random forest algorithm to calculate the importances of attributes in ten

FIGURE 4: The relation between local sensitivity and threshold ϵ .

FIGURE 5: Clustering results on artificial datasets based on HDBSCAN. (a) Initial division for cluto-t5-8k. (b) Aggregation for cluto-t5-8k. (c) Initial division for cluto-t8-8k. (d) Aggregation for cluto-t8-8k. (e) Initial division for triangle2. (f) Aggregation for triangle2.

clusters and thus analyze the characteristics of each cluster. Specifically, based on the labels generated by HDBSCAN clustering, we take each cluster as the positive class and the other clusters as the negative class to construct multiple binary classification models, thereby mining the important attributes within each cluster.

Based on the important attributes within clusters, the characteristics of Cluster 1 are listed below. (1) The household has no children under the age of 16. (2) The annual net income of the household is higher than the average level. (3) Medical expenses are more prominent in the expenditure of the household. The characteristics of

TABLE 3: The metric values on four UCI datasets.

Dataset	Metric	AHC	DBSCAN	EDBSCAN	NS-DBSCAN	ADBSCAN	HDBSCAN
Banknote	SC	0.308	0.321	0.481	0.483	0.484	0.485
	DBI	1.239	1.492	1.226	1.197	1.875	1.155
	ARI	0.796	0.753	0.865	0.912	0.987	0.991
	NMI	0.758	0.865	0.897	0.953	0.992	0.996
Planning relax	SC	0.178	0.136	0.171	0.215	0.231	0.271
	DBI	2.355	11.834	1.733	2.322	2.095	2.368
	ARI	0.612	0.537	0.712	0.739	0.865	0.898
	NMI	0.821	0.854	0.882	0.913	0.924	0.928
Parkinson	SC	0.258	0.201	0.212	0.255	0.276	0.281
	DBI	1.731	1.785	1.598	1.679	1.717	1.640
	ARI	0.635	0.689	0.728	0.825	0.877	0.969
	NMI	0.589	0.721	0.737	0.862	0.883	0.928
Codon usage	SC	0.265	0.238	0.275	0.281	0.283	0.296
	DBI	3.856	8.954	2.917	2.805	2.655	2.192
	ARI	0.712	0.884	0.865	0.928	0.944	0.939
	NMI	0.822	0.851	0.874	0.912	0.953	0.967
HCV	SC	0.328	0.337	0.416	0.397	0.419	0.511
	DBI	2.754	2.663	2.841	2.425	2.538	2.331
	ARI	0.714	0.821	0.885	0.836	0.918	0.986
	NMI	0.689	0.774	0.796	0.825	0.884	0.917

TABLE 4: The metric values on the CFPS2016 datasets.

Metric	AHC	DBSCAN	EDBSCAN	NS-DBSCAN	ADBSCAN	HDBSCAN
SC	0.099	0.109	0.138	0.158	0.177	0.237
DBI	2.656	2.019	1.977	1.765	1.528	1.003
ARI	0.713	0.852	0.872	0.893	0.948	0.995
NMI	0.788	0.854	0.861	0.875	0.942	0.993

TABLE 5: The accuracy values on the CFPS2016 datasets.

Metric	AHC	DBSCAN	EDBSCAN	NS-DBSCAN	ADBSCAN	HDBSCAN
Accuracy	79.6%	82.7%	85.3%	85.8%	87.1%	89.6%

TABLE 6: The result of clustering based on HDBSCAN.

Label	1	2	3	4	5	6
Number	382	369	282	148	127	126
Label	7	8	9	10	Noise	
Number	90	89	61	34	70	

Cluster 9 are as follows: (1) the average age of adults in the household is 76. (2) Almost every household member has no pension insurance. Besides, the characteristics of Cluster 10 are as follows: (1) the annual per capita income of the household is 35,914 yuan, 1.43 times higher than the average level. (2) More than half of the members use computers. The living standard of households in Cluster 10 is relatively high compared with other clusters, and Cluster 10 accounts for a small proportion of poverty-stricken households. According to the above analysis, the causes of poverty and characteristics for most households are similar so that the numbers of households in some clusters are large whereas the characteristics of a few poverty-stricken households are clearly different from others, which leads to small numbers of households in clusters such as Cluster 9 and Cluster 10.

Figure 6 shows the distribution of attribute importances in each cluster, where the abscissa values indicate the numbers of 320 attributes and the ordinate values indicate the attribute importances; the ten curves represent the distribution of attribute importances in ten clusters.

As can be seen from Figure 6, the distributions of attribute importances represented by ten curves nearly differ from each other. For instance, the attribute with the highest importance in Cluster 7 is the 165th-dimensional attribute which denotes the stage of schooling for household members at the last survey. And that in Cluster 8 is the 218th-dimensional attribute which denotes the total post-tax annual income from work. The phenomenon shows that poverty-stricken households within different categories differ in the characteristics and the causes of

FIGURE 6: The attribute importances based on the clustering result by HDBSCAN.

TABLE 7: The metric values on four UCI datasets.

Dataset	AHC	DBSCAN	EDBSCAN	NS-DBSCAN	ADBSCAN	HDBSCAN
Banknote	2.7	2.6	2.8	2.9	2.8	3.1
Planning relax	2.3	2.1	2.5	2.6	2.6	2.3
Parkinson	2.5	2.4	2.7	2.6	2.7	2.5
Codon usage	2.9	2.7	2.9	2.8	2.9	2.8
HCV	2.6	2.4	2.7	2.8	2.6	2.5
CFPS2016	2.8	2.4	2.6	2.7	2.7	2.5

poverty. Therefore, the proposed method can identify the commonalities and differences in poverty effectively. Finally, for all the datasets, we conduct computational complexity experiments with the different methods. The results are shown in Table 7. Because the proposed method is the hierarchical DBSCAN algorithm based on the initial division and aggregation of neighbor clusters, the time is higher than traditional DBSCAN. However, the time is lower than other new methods.

5. Conclusions

This paper designs the hierarchical DBSCAN algorithm based on the initial division and aggregation of neighbor clusters. First, the proposed method HDBSCAN adopts the adaptive neighborhood radius to perceive regions with different densities and thus makes the initial division of the dataset. Then, iterative aggregation is performed on neighbor clusters according to the border and inner distances. Experiments on artificial datasets and UCI real-world datasets indicate that HDBSCAN has the ideal performance of clustering. Additionally, HDBSCAN divides the dataset of Chinese poverty-stricken household, namely, CFPS2016, into 10 clusters, and experimental results verify the rationality of the clustering result. The main reasons for the ideal performance of HDBSCAN lie in

the following two aspects. First, the adaptive neighborhood radius helps to identify regions of different densities in the data space with imbalanced density distribution. Second, the aggregation further merges neighbor clusters with similar densities, which weakens the impact of the accuracy of initial partition on the clustering performance effectively. However, if the dimension of the datasets is very higher, the cluster effect is not better. In the future, more research studies will be conducted on the clustering result of the CFPS2016 dataset. To be specific, we will study the characteristics of poverty-stricken households in each category so as to support the formulation and implementation of antipoverty measures. The advanced clustering technology will be applied in targeted poverty alleviation of the poverty counties in China.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

- [1] H. Hsieh, S. Lin, and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data," in *Proceedings of the 21th ACM SIGKDD International Conference*, Sydney, Australia, August 2015.
- [2] P. Li, Z. Chen, L. T. Yang, Q. Zhang, and M. J. Deen, "Deep convolutional computation model for feature learning on big data in internet of things," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 790–798, 2018.
- [3] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "Incremental deep computation model for wireless big data feature learning," *IEEE Transactions on Big Data*, vol. 6, no. 2, p. 248, 2020.
- [4] Z. Lv, L. Qiao, K. Cai, and Q. Wang, "Big data analysis technology for electric vehicle networks in smart cities," *IEEE Transactions on Intelligent Transportation Systems*, p. 1, 2020.
- [5] J. Chen, Z. Lv, and H. Song, "Design of personnel big data management system based on blockchain," *Future Generation Computer Systems*, vol. 101, pp. 1122–1129, 2019.
- [6] Z. Lv, X. Li, H. Lv, and W. Xiu, "BIM big data storage in WebVRGIS," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2566–2573, 2020.
- [7] J. E. Blumenstock, "Fighting poverty with data," *Science*, vol. 353, no. 6301, pp. 753–754, 2016.
- [8] Y. Zhou, Y. Guo, Y. Liu, W. Wu, and Y. Li, "Targeted poverty alleviation and land policy innovation: some practice and policy implications from China," *Land Use Policy*, vol. 74, pp. 53–65, 2018.
- [9] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Vol. 344, John Wiley & Sons, Hoboken, NJ, USA, 2009.
- [10] Y. Xie and J. Hu, "An introduction to the China family panel studies (CFPS)," *Chinese Sociological Review*, vol. 47, no. 1, pp. 3–29, 2014.
- [11] A. Saxena, M. Prasad, A. Gupta, N. Bharill, and O. P. Patel, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.
- [12] H. Tiwari, "Clustering algorithm and its application in data mining," *Wireless Personal Communications*, vol. 110, no. 1, pp. 21–30, 2020.
- [13] V. Cohen-Addad, V. Kanade, F. Mallmann-Trenn, and C. Mathieu, "Hierarchical clustering," *Journal of the ACM*, vol. 66, no. 4, pp. 1–42, 2019.
- [14] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "DBSCAN: past, present and future," in *Proceedings of the Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT)*, pp. 232–238, Chennai, India, February 2014.
- [15] X. Yu, D. Zhou, and Y. Zhou, "A new clustering algorithm based on distance and density," in *Proceedings of 2005 International Conference on Services Systems and Services Management*, vol. 2, pp. 1016–1021, Chongqing, China, June 2005.
- [16] A. Ram, S. Jalal, A. S. Jalal, and M. Kumar, "A density based algorithm for discovering density varied clusters in large spatial databases," *International Journal of Computer Applications*, vol. 3, no. 6, pp. 1–4, 2010.
- [17] P. Liu, D. Zhou, and N. Wu, "VDBSCAN: varied density based spatial clustering of applications with noise," in *Proceedings of IEEE International Conference on Service Systems and Service Management*, pp. 1–4, Chengdu, China, June 2007.
- [18] A. Fahim, A.-E. Salem, F. Torkey, M. Ramadan, and G. Saake, "Scalable varied density clustering algorithm for large datasets," *Journal of Software Engineering and Applications*, vol. 3, no. 6, pp. 593–602, 2010.
- [19] M. N. Gaonkar and K. Sawant, "AutoEpsDBSCAN: DBSCAN with Eps automatic for large dataset," *International Journal on Advanced Computer Theory and Engineering*, vol. 2, no. 2, pp. 11–16, 2013.
- [20] A. Ram, A. Sharma, A. S. Jalal, A. Agrawal, and R. Singh, "An enhanced density based spatial clustering of applications with noise," in *Proceedings of the 2009 IEEE International Advance Computing Conference*, pp. 1475–1478, Patiala, India, March 2009.
- [21] A. Kouhi, H. Seyedarabi, and A. Aghagolzadeh, "Robust FCM clustering algorithm with combined spatial constraint and membership matrix local information for brain MRI segmentation," *Expert Systems with Applications*, vol. 146, 2019.
- [22] D. Fan, Z. Song, S. Jon et al., "An improved quantum clustering algorithm with weighted distance based on PSO and research on the prediction of electrical power demand," *Journal of Intelligent and Fuzzy Systems*, vol. 38, no. 2, pp. 1–9, 2020.
- [23] M. Xu, J. Zhou, and P. Zhu, "An electronic nose system for the monitoring of water cane shoots quality with swarm clustering algorithm," *Journal of Food Safety*, vol. 41, no. 1, Article ID e12860, 2020.
- [24] Y. Chen, L. Zhou, N. Bouguila et al., "BLOCK-DBSCAN: fast clustering for large scale data," *Pattern Recognition*, vol. 109, 2021.
- [25] S. Chel, S. Gare, and L. Giri, "Detection of specific templates in calcium spiking in HeLa cells using hierarchical DBSCAN: clustering and visualization of CellDrug interaction at multiple doses*," in *Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2425–2428, Montreal, Canada, July 2020.
- [26] F. Znidi, H. Davarikia, M. Arani, and M. Barati, "Coherency detection and network partitioning based on hierarchical DBSCAN," in *Proceedings of the 2020 IEEE Texas Power and Energy Conference (TPEC)*, pp. 1–5, College Station, TX, USA, August 2020.
- [27] M. Parmar, D. Wang, X. Zhang et al., "REDPC: a residual error-based density peak clustering algorithm," *Neurocomputing*, vol. 348, pp. 82–96, 2019.
- [28] 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.
- [29] 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, November 2017.
- [30] G. E. Batista and M. C. Monard, "A study of k-nearest neighbour as an imputation method," *Second International Conference on Hybrid Intelligent Systems*, vol. 87, pp. 251–260, 2002.
- [31] S. Alkire and J. Foster, "Understandings and misunderstandings of multidimensional poverty measurement," *The Journal of Economic Inequality*, vol. 9, no. 2, pp. 289–314, 2011.
- [32] P. J. Rousseeuw and "Silhouettes," *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, *Journal of Computational and Applied Mathematics*, vol. 20, no. 1, pp. 53–65, 1987.
- [33] A. Bryant and K. Cios, "RNN-DBSCAN: a density-based clustering algorithm using reverse nearest neighbor density

- estimates,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1109–1121, 2018.
- [34] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.
- [35] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proceeding of 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, Menlo Park, CA, USA, June 1996.
- [36] T. Wang, C. Ren, Y. Luo, and J. Tian, “NS-DBSCAN: a density-based clustering algorithm in network space,” *ISPRS International Journal of Geo-Information*, vol. 8, no. 5, p. 218, 2019.
- [37] H. Li, X. Liu, T. Li et al., “A novel density-based clustering algorithm using nearest neighbor graph,” *Pattern Recognition*, vol. 102, 2020.