IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 24, 2019, accepted June 11, 2019, date of publication June 19, 2019, date of current version July 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923694

Improved Density Peak Clustering Based on Information Entropy for Ancient Character Images

YU WENG[1], **NING ZHANG**[1], **AND XIAOXIAN YANG**[2]

[1]School of Information Engineering, Minzu University of China, Beijing 100081, China
[2]School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai 201209, China

Corresponding author: Ning Zhang (nzhang1995@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772575, in part by the National Key R&D Program of China under Grant 2017YFB1402101, in part by the MUC 111 Project, in part by the Youth Foundation of Shanghai Polytechnic University under Grant EGD18XQD01, and in part by the CERNET Innovation Project under Grant NGII20170513.

ABSTRACT A large number of IoT applications require the use of supervised machine learning, a type of machine learning algorithm that requires data to be labeled before the model can be trained. Because manually labeling large datasets is a time-consuming, error-prone, and expensive task, automated machine learning methods can be used. To tackle the challenge in which an ancient character image needs to be manually labeled, this paper explores the classification method of ancient Chinese character images based on density peak clustering. We design a metric function of density peak clustering and propose an improved density peak clustering method based on information entropy for ancient book image classification. The method enumerates the distance threshold of clustering, then calculates the information entropy of the clustering result, and determines the class distance threshold by analyzing the attenuation of the information entropy, thereby completing the image clustering process. The improved metric function is used to calculate the similarity between images. A greedy strategy is used as the basis of the merging operation of the class members to achieve the purpose of increasing the degree of information entropy attenuation. The experimental results on the dataset of the Yi character images prove that the method can accurately classify unknown character images of ancient books.

INDEX TERMS Density peak clustering, information entropy, ancient character image.

I. INTRODUCTION

Machine learning in the IoT brings an interesting problem: the best models need to be trained on large amounts of data, and most IoT devices are still limited by storage space and processing power. Therefore, the ability to efficiently transfer large amounts of data from IoT devices to servers or the cloud and improve data output is key to smart application development. On mobile devices, numerous services offer similar functionalities, and therefore it is important to select an appropriate choice among these services [1]. An ensemble learning-based method was proposed in [2] that is able to identify similar neighbors and filter out fake neighbors. The IoT generates huge amounts of data, which is one of its defining characteristics. To manually label these data is a time-consuming, error-prone, and expensive task, and therefore machine learning methods are preferred. IoT devices generate large amounts of heterogeneous data such as images, audio, language, and time series, and different automated methods are required to process these data. Image classification technology is mainly used for optical character recognition, handwritten text recognition, and the management and retrieval of ancient character image data, and it is very important in research on the recognition, retrieval and labeling of ancient character images. In image classification studies, supervised classification often requires labeled data. Usually, in the process of studying a model, the modeled training and testing will be carried out using labeled public datasets. In research on the image recognition and retrieval of ancient books, the image dataset generally needs to be normalized into the data form corresponding to the model. These datasets often must be manually labeled and processed, which takes a

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

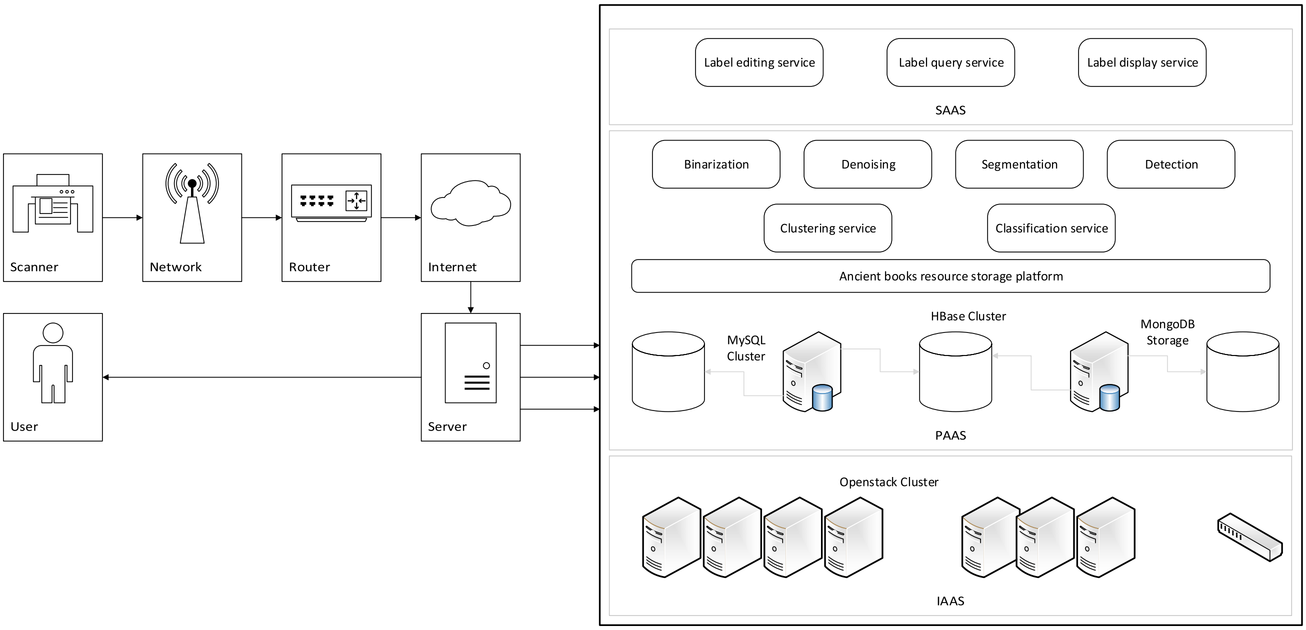


FIGURE 1. System service architecture.

lot of manpower and material resources. These manpower and time requirements additionally hinder the research progress of the subject. Since people have different criteria for judging things, manual labels may be inconsistent, and there may be a risk of incorrect labeling during labeling work that takes place over a long period of time. In summary, the consistency and objectivity of manual labeling are not guaranteed. Using samples with incorrect annotations can have negative effects, such as reduced image recognition rates and image retrieval errors. Therefore, it is especially important to study and explore how to better label ancient images. The system service architecture is shown in Fig. 1. To select or recommend useful services, many types of information can be utilized to improve the results, such as location [3], [4], [5]. The behavior of the system reflects its functions, so the design of a service system is an important process [6]. The evaluation of a service [7] can help to improve its quality. Researchers have been studying clustering analysis techniques for a long time, and a large number of clustering algorithms have emerged. According to their core ideas, they can be divided into partition-based clustering methods, hierarchical-based clustering methods, spectral clustering methods, and density-based clustering [8], [9]. Classical clustering algorithms based on partitioning ideas include algorithms such as K-means and K-medoid [10], which are complicated by the calculation of the distance between all sample points after each adjustment of the grouping of one sample point. The calculation increases exponentially with the increasing amount of data, and therefore it is not suitable for the cluster analysis of large-scale data. Such clustering algorithms often need to give the number of target clusters and cannot achieve the clustering of nonspherical data well. In practical applications, such algorithms tend to be completely different because of the given key thresholds of clustering, and it is possible to obtain diametrically opposite clustering results. Hierarchical-based clustering algorithms usually have two clustering strategies: bottom-up and top-down. The condensed hierarchical clustering algorithm is a classical hierarchical clustering algorithm. However, the hierarchical clustering method cannot adjust the upper layer structure in the process of clustering, and its time complexity is at least the square of the data amount, which makes it inefficient when clustering large-scale data. The hierarchical clustering algorithm has a high dependence on the critical threshold of clustering, and under different thresholds, it will produce completely different results. Based on graph theory and matrix analysis, the method of spectral clustering is used to map high-dimensional data in low-dimensional space by calculating the eigenvalues and eigenvectors of the Laplacian matrix, which can effectively process irregular data [11]. The density-based clustering method replaces the similarity between sample points by the density of sample points in space. According to the spatial density distribution of sample points, the spatially large enough spaces are grouped together to form a cluster. This method is capable of finding a family of any shape in the sample space and is able to eliminate noise to some extent. Classical density clustering methods include DBSCAN [12], OPTICS [13], and DENCLUE [14]. In recent years, Rodriguez and Laio proposed a density peak clustering algorithm [15]–[17]. This method is very popular because of its simple idea and good clustering effect [14], [16], and it has also been widely applied [18]. Aiming at the obvious characteristics of the feature image of ancient Chinese text segmentation, we propose an improved method based on density clustering for already segmented good text images, using the information entropy to evaluate the optimal density clustering value. It is able to accurately classify unknown ancient document character images. The main contributions of this work are as follows:

- We constructed a handwritten Yi character dataset. The Yi characters used in this paper are cut from the scanning documents of Yi classics, and we apply normalization and binarization to these image slices. After these steps, the images we obtained are binary images of the size 100 × 100.
- We propose an improved density peak clustering algorithm that uses the change of entropy to find the best clustering threshold. It is effective under the situation of large difference of density.

The rest of this paper is organized as follows. We first provide a brief review of related work in Section 2. Then, we describe our proposed method in Section 3. Before concluding, we present our experimental results in Section 4.

II. RELATED WORK

In the IoT environment, various devices generate different types of data, and it is necessary to analyze the hidden complex features behind these data. Yin et al. used a matrix factorization integrated convolutional neural network (CNN) to learn deep features [19], and on service recommendation tasks, their method achieves a high quality-of-service (QOS) prediction accuracy. In general, the image annotation can be implemented either manually or automatically. ImageNet is a classic image database for manually annotating datasets [20]. In addition, its annotation work is mainly achieved by the ''Amazon Mechanical Turk'' platform based on crowdsourcing technology utilizing 48,940 staff from 167 countries [21]. This scale of the manual annotation of image methods is not affordable to the average organization or individual, as it consumes a lot of manpower, material and financial resources. However, many researchers who have manually labeled datasets have also designed many image annotation tools, such as MIT's LabelMe [22] and IAT-Image Annotation Tool [23]. Some commonly used tools are shown in Table 1. Mature results and technologies have been achieved for the problem of the implementation of automated image organization and annotation by distributing known metadata information. In the traditional image annotation method, feature design and representation are first carried out, including image preprocessing, feature extraction and feature selection, which are also commonly referred to as feature design and representation. The process mainly extracts features for the image color, image texture, geometric shape, light and dark light and then uses a machine learning algorithm to classify and divide and thereby complete the labeling process [24], [25]. In the automation research of image annotation, there are classification-based image annotation methods, image annotation methods based on an overview association, and search-based annotation methods. Cusanod *et al.* proposed a support vector machine method characterized by a color histogram [26]. In the same period, Chang *et al.* proposed a Bayesian point learning machine method based on color texture features [27]. Li and Wang proposed a hidden Markov method based on the wavelet transform [28]. Additionally, Yang *et al.* proposed a multi-instance learning method based on image segmentation and regional underlying blending features [29]. These studies are all classification-based annotation models. These methods obtain relatively good results, but the model is not generalized, still depending on the annotation of the initial sample, and is complicated. The research on and development of classification-based annotation models is closely linked to the development of machine learning classification techniques. Comaniciu and Meer [30] proposed the Mean shift algorithm, which is a commonly used local feature extraction method and can be used to implement classification annotation. There are also visual word package models [31], Fisher vectors [32], SPM methods based on sparse coding, locality-constrained linear coding, and SVMs [26], [33]. With the development of deep learning, model-based image annotation technology has made a great breakthrough. Vinyals *et al.* proposed an automatic image annotation based on the Encoder-Decoder framework. The method proposes features by a convolutional neural network and then applies LSTM to generate the target language, thus generating picture description text [34]. Xu *et al.* proposed two different attention mechanisms to extract image attributes through multi-instance learning for label text reorganization and complete image description [35]. Zhou *et al.* used the Tex-Conditional method to combine image features to complete the annotation of images [36]. Wang *et al.* proposed a method based on CNN-RNN [37] that uses the VGGNet model to perform training and then multitag training and finally recognizes the detection. The input LSTM is used to describe the generation [38]. This method preserves the high-level semantic information of the image. At present, most of the methods of image annotation using deep learning are based on improving the structure and parameters of neural networks and improving the accuracy of the labeling [39]. This type of method is mainly based on neural network structures such as CNN, RNN, SSD, YOLO, and FastRNN [37], [40]. In the image annotation method based on probability association, the correlation between the image region feature and the labeled word is mainly analyzed, and then the new image annotation word is inferred. Mori *et al.* proposed a method for labeling images by using keywords and a visual vocabulary. This is the earliest proposed image annotation [39]. Bouzaieni and Tabbone proposed an image annotation method based on the overview graph model [41]. Search-based annotation enables the automatic annotation of images through massive image retrieval and an analysis of the network [42], [43]. With the research on and development of the Semantic Web, many research results have been achieved through the Semantic Web-based analysis of annotation technology [44], [45]. In summary, the problem of labeling a sample set remains unresolved for images of ancient books. Therefore, in order to address these challenges, we propose an image labeling method based on density peak clustering. This method not only can complete the unsupervised annotation of images

TABLE 1. Descriptions of common tools for image annotation.

Tools	Description
Pixorize	Users can upload, tag, and share images.
Labelbox	A data-tagging platform for training expert machine learning applications. Support for JSON / CSV /WKT / COCO / Pascal VOC export data.
FastAnnotation SingleObject	Can be used to quickly label a large number of images (draw bounding boxes and assign labels) in PASCAL VOC format.
Microsoft VoTT	A cross-platform tool for tagging videos and images. Support for the integration of CNTK, TensorFlow and YOLO in a deep learning framework and support for computer-aided object tracking.
Annotorious	An open-source image annotation toolkit written in JavaScript.
FastAnnotationTool	Tools for image annotation using OpenCV can be applied to image classification, optical character reading, and more.
LabelImg	A graphical image annotation tool that can perform label object bounding box operations.
LabelMe	An online annotation tool developed with the goal of building an image database for computer vision research.
Lear Image Annotation Tool	A tool that uses a bounding box to annotate objects in an image.
Ratsnake Image Annotation Tool	Allows manual labeling by using polygons, splines, and meshes. A customizable active contour model is integrated to achieve the semiautomatic segmentation of objects.
Jsoda	A simple JavaScript web application for bounding box annotations for object detection.
Philosys Label Editor	Allows full annotation of 2D images and 3D scenes such as point clouds with geometric markers, attributes and semantic segmentation.

but can also improve the degree of image automation by combination with the classification-based image annotation method, the image annotation method based on the introduction, or the search-based annotation method.

III. METHODS

Since the DPC algorithm needs to map the product χ and ρ according to the class density ρ and the minimum interclass distance δ, the class center threshold is manually selected. Therefore, as the distribution of the relationship map is different, the results of the manual selection will be different. When the amount of data of the members is particularly large, it is very difficult to construct a relationship diagram of all the member points and manually select them. Therefore, this paper proposes a density peak clustering method based on an information entropy improvement. The core idea of this method is to use the distance threshold to enumerate the possible values and perform clustering to obtain an iterative solution for the information entropy. The experimental results show that upon merging classes, the information entropy will undergo fault attenuation. By determining the number of faults, the number of suspicious clusters can be estimated. The use of greedy strategies and the sharing of nearest neighbors for the expansion of intraclass elements can make the attenuation of the information entropy more obvious.

A. CLUSTERING FEATURE SELECTION AND SIMILARITY MEASURE

1) MINKOWSKI DISTANCE

To complete the merge operation of the set in the target space of the cluster, the similarity calculation of the target object is required. Commonly used similarity calculation functions have a Markov distance and a Mahalanobis distance. In the given clustering target space, any two pictures obey the same distribution and for any picture i, there are $X_i = (x_i^1, x_i^2, x_i^3, \cdots)$. m, n are the row and column values of the picture, respectively. The Minkowski distance is defined as (1), where $i > 0, j > 0, k > 0, r \geq 1$, and Σ_{ij} is the covariance matrix corresponding to X_i and X_j.

$$d(X_i, X_j) = \left(\sum_{k=1}^{n} \|x_k - y_k\|^r\right)^{\frac{1}{r}} \tag{1}$$

2) MAHALANOBIS DISTANCE

The Mahalanobis distance is the covariance distance of the data. It can calculate the similarity between two samples more effectively. Different Euclidean distances are scale-independent. The Mahalanobis distance is defined as (2). If the covariance matrix is a unit matrix or a diagonal matrix, the Mahalanobis distance is equal to the Euclidean distance.

$$d(X_i, X_j) = \sqrt{(X_i - X_j)^T \Sigma^{-1} (X_i - X_j)} \tag{2}$$

3) SELF-DEFINED SIMILARITY

According to the key point distribution characteristics of ancient book image data, a self-defined metric can be used to calculate the similarity between two images. To comprehensively consider the data distributed in the cluster space, we defined a set of equations as (3)-(6). Specifically, the description is given as follows: let M be the point quality of the image, which represents the information entropy value in the two-dimensional space of the given image. ρ is the point density, which represents the information entropy value of each key point calculated. F is the point force used to evaluate the tightness between multiple images, and J is the Mahalanobis distance between the two images.

$$M = -\sum_{i=n}^{L} \frac{x_i}{m*n} \log\left(\frac{x_i}{\sum_{j=0}^{m*n} x_j}\right) \tag{3}$$

$$\rho = \frac{M}{\sum_i^L l_i} \tag{4}$$

$$J = \sqrt{(X_i - X_j)^T \Sigma^{-1} (X_i - X_j)} \tag{5}$$

$$F = \frac{M_1 * M_2}{J^2} \tag{6}$$

B. DENSITY PEAK CLUSTERING (DPC)

In the DBSCAN algorithm, much effort is required to determine the two thresholds of *MinPts* and *Eps*, and the threshold is completely dependent on experimental tuning and research experience. The parameter *Eps* defines the radius of the neighborhood around a point x; it is called the ϵ-neighborhood of x. The parameter *MinPts* is the minimum number of neighbors within the "*Eps*" radius. Compared with the DBSCAN algorithm, in the DPC algorithm, we first calculate the density value ρ_i of the data points and the minimum interclass distance δ_i of the data points and then calculate the product γ of the data point density values ρ_i and δ_i. It is possible to construct a map of the δ and γ distributions of the members of the collection so that the center of the class can be visually discovered. In the DPC algorithm, the calculation formula of the density ρ is defined as (7). D_c is the clustering threshold of the model, which is the number of members used to classify the class. In [15], it is mentioned that δ_i is calculated according to the formula (8), which is the minimum value of d_{ij} that satisfies ρ_j greater than ρ_i from point i to all sample points, which can divide the points into different centers. The specific formulas for ρ_i and δ_i are given as (7) and (8), respectively.

$$\rho_i = \sum \chi(d_{ij} - d_c) \tag{7}$$

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{8}$$

For ρ_i, let $x = d_{ij} - d_c$, when $x < 0$, then $\chi(x) = 1$; otherwise, $\chi(x) = 0$. Therefore, $\sum \chi(x)$ represents the sum of the points i to all sample points satisfying d_{ij} less than d_c. In [16], the Gaussian kernel function is used for dataset processing, and the processed result represents the local density of the data field at point i. In the experiment, we also find that as the dataset size increases, the calculation cost of parameters such as ρ_i will increase. To address this issue, we design δ with local dynamics. On the other hand, for the algorithm, the maximum value δ_i of d_{ij} satisfying ρ_j greater than ρ_i from point i to all sample points can be obtained as (9).

$$\delta_i = \max_{j:\rho_j > \rho_i} (d_{ij}) \tag{9}$$

C. IMPROVE DPC WITH INFORMATION ENTROPY

In the clustering method based on the density peak, although the class center can be quickly found by hand, the class division of the set is determined. However, this method has certain constraints on the applicable scene. For example, when the amount of data is huge, it is difficult to practice the method of constructing the distribution map of ρ and δ to obtain the threshold. Moreover, since it is a distribution map for γ and δ, there is a distribution condition that is not easily distinguishable, which introduces difficulty in manually selecting a threshold. To address the latter problem, new variables can be constructed to reduce the likelihood of distribution symmetry.

1) SELECTION OF CLASS DENSITY CENTER

In DPC, the center of the class density is selected according to the class density ρ_i, and it is solved by the formula (7). However, in [15], the center of the γ decision class is manually evaluated and depends on the product of ρ_i and δ_i. In the model of this paper, no manual intervention is needed to evaluate the class center. For each different d_c value, the corresponding class density center will change. Once the class density center is determined, a partition-based clustering method can be used. This paper chooses to use *shared nearest neighbor* (SNN) [46] for clustering.

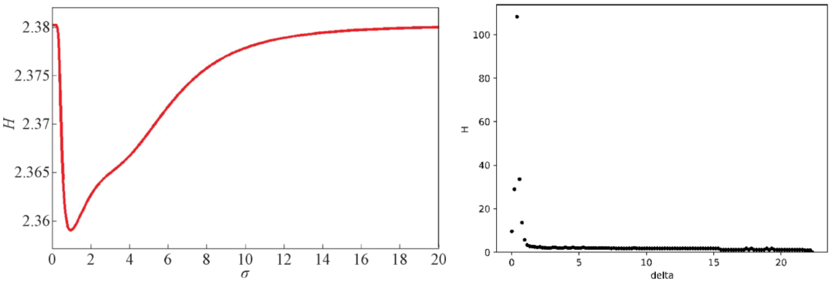
2) STEP SIZE

In [16], the starting point for the growth of d_c is from 0 to infinity. To speed up the enumeration speed of d_c, we use the variance of the metric as a growth factor to control the growth rate, which can speed up the enumeration speed. If the drop in entropy is found to be severe, then d_c is backtracked, and the pace of growth is slowed down to d_c percent, traversing the enumeration, and similar growth operations are repeated until the d_c enumeration is complete or H is attenuated to 0. The variance can represent the degree of dispersion in the similarity measure. The standard deviation (10) can be used as the growth step to describe the average of the distances from each sample point to the mean. When the information entropy drops sharply under the standard deviation, the current pace of d_c is used as step size, and enumerating from the last moment, it can speed up the enumeration and avoid the loss of important d_c values.

$$S = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \bar{D})^2}{n-1}} \tag{10}$$

3) INFORMATION ENTROPY EVALUATION

In the DPC algorithm, when the number of elements in the range of d_c is larger, it means that the information value existing in the d_c range is higher. Then, if the elements in the range can be regarded as a class, then when the class merge operation occurs, the information entropy will drop rapidly. Because the entire space becomes more consistent and less confusing, the value of the information is also reduced. As classes merge, the information entropy will eventually equal zero. To this end, we propose to use information entropy to improve the DPC method. According to the local density ρ of each point and the minimum distance value d,



(a) (b)

FIGURE 2. The entropy changes with the growth of δ.

the point where the local density is the largest and that has the largest distance value is selected as the cluster center, and then the clustering is completed. When d_c takes different values, the local density ρ and the minimum distance value d of different sample points are obtained, and at this time, a corresponding measure of the entropy is obtained in the model. By evaluating the falling gradient value of the information entropy, we can know that the value of d_c obtains its optimal density clustering result at the stagnation point. The information entropy is calculated by (11).

$$H = -\sum_{i=1}^{n} \frac{c_i}{N} \log\left(\frac{c_i}{N}\right) \tag{11}$$

In (11), C_i is the number of members of the first cluster and N is the total number of sample points. When the initial time of the cluster is d_c, then each member in the system model is a separate class, and then the information entropy H has its maximum value. As d_c increases gradually, the information entropy value of the system will gradually decrease, the degree of change of the entropy value of the system under different d_c values will be analyzed, and the optimal value of d_c can be finally determined. The proposed solution is to use the information entropy stability range to evaluate the aggregation state of the cluster without using γ_i. In this algorithm, the values of δ_i are subjected to a process of cluster execution, and the feasibility of calculating and evaluating the new data point merging operation is dynamically determined. The specific algorithm is shown in Algorithm 1.

Algorithm 1 Calculate Distance Threshold d_c

```
Input: dataset $X = (x_1, x_2, \cdots, x_n)$
Output: $d_c$
Distance $d_{ij} = distance(x_i, x_j)$
$d_c = \min_j(\min_i(d_{ij}))$, $d_{max} = \max_j(\max_i(d_{ij}))$
Step size: $Learning_{step} = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}$
repeat
    $\rho_i = \sum \chi(d_{ij} - d_c)$
    $pile_i = \{i | \forall i, \rho_i = \max_i(\rho)\}$
    Sort $\rho_i$ in descending order
    $pile_i = \{i | \exists i, \rho_i = \max(\rho)\}$
    initialize $pile_i = \{pile_1\}$
    repeat
        $pile_i = \{k | k \notin pile$, and $\forall k, \rho_k = \max(\rho)\}$
    until $\rho_i = \{\phi\}$
    Merge class $pile$
    $H = -\sum_{i=1}^{n} \frac{c_i}{n} \log\left(\frac{c_i}{n}\right)$
until The value of $H$ converges
return $d_c$
```

TABLE 2. Dataset description.

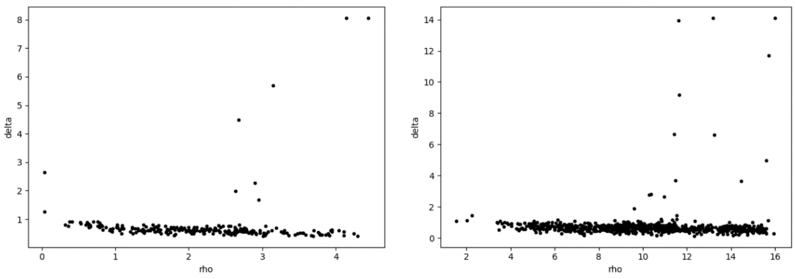
Dataset	# Labels	# Instances
Aggregation	7	788
Flame	2	240
Spiral	3	312
Unbalance	8	6,500

IV. EXPERIMENTS

A. PUBLIC DATASET

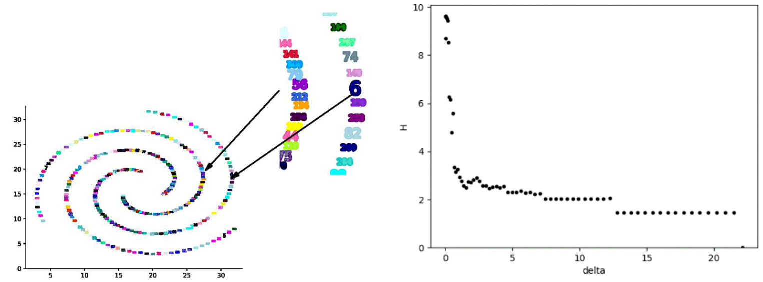
We select several public datasets with which to conduct experiments, as described in [15], including Aggregation, Flame, and Unbalance. Detailed descriptions of these datasets are shown in Table 2. Using Algorithm 1, we obtain the experimental results. For the Aggregation dataset, Fig. 2(a) shows the results in [16]. The comparison shows that the threshold can be found more clearly by a reasonable Gaussian density (mainly by finding the inflection point), but this method relies heavily on the selection of kernel functions, and the interclass Gaussianization may produce new errors. According to Fig. 2(b), the distance parameter Eps, and the class evaluation threshold $MinPts$, the density peak clustering is performed. The ρ and δ relationships calculated by Flame and Aggregation based on the Gaussian kernel function are shown in Fig. 3. Although the range of values selected by Flame's data center can be clearly seen in the left figure, for Aggregation, the selection of cluster center points is not easy.

However, the possible prediction values generated by the information entropy enumeration evaluation method can be used to find the exact range of threshold values. For example, in Fig. 4, (b) is the convergence of the suspicious threshold of the path dataset. In Fig. 4(b), the manual selection is the observation result when the d_c value is 0.05. At this time, the number of clusters is large, the system is unstable, and the information entropy is large. With the further d_c growth,



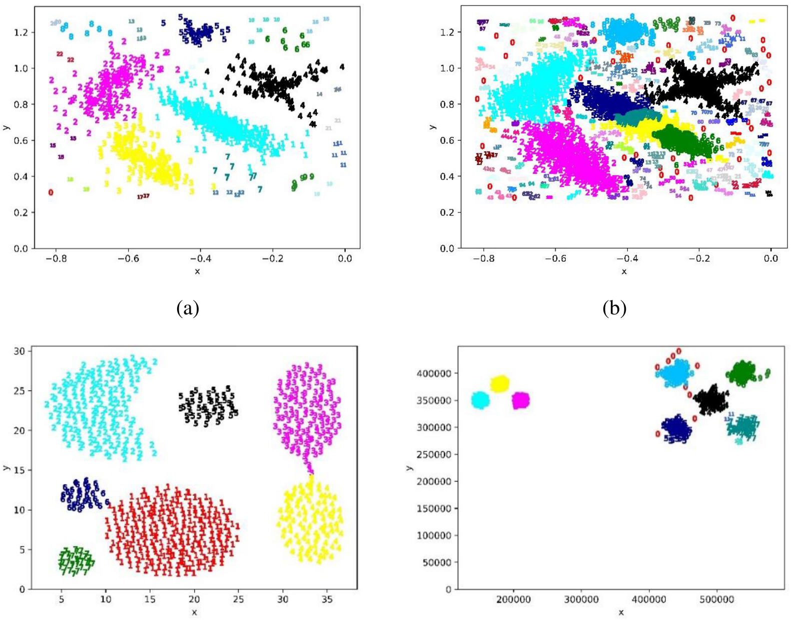
(a) (b)

FIGURE 3. Threshold distributions of flame and aggregation.



(a) (b)

FIGURE 4. Clustering for the path dataset.



(a) (b)

(c) (d)

FIGURE 5. Clustering results.

the number of classes in the system will decrease; after falling to a stable state, the d_c growth will only merge when it reaches the minimum distance between classes, and the information entropy will decrease. Prior to this, the information entropy was stable. Therefore, it can be clearly seen from Fig. 4(b) that d_c will have hierarchical information entropy attenuation. Fig. 5 shows the effect of clustering on the above-mentioned datasets. According to the experimental results, the improved density peak clustering algorithm solves the problem that a manual input threshold is required.

B. YI CHARACTER IMAGE DATASET

The Yi people are a minority in China, and the words they use are called the Yi language. The writing format is straight from the left, and there are no punctuation marks. The literature is rich in content and covers all aspects of the ancient

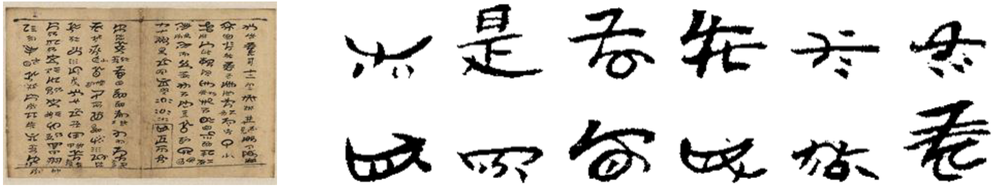


FIGURE 6. Yi character image examples.

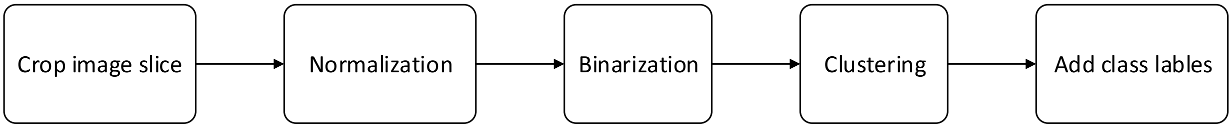


FIGURE 7. Construction procedure of the dataset.

TABLE 3. Accuracy.

Method	Accuracy
K-means	78 ±0.4%
DBSACN	85 ±0.3%
DPC	87 ±0.3%
Improved DPC	91 ±0.5%

society of the Yi people, objectively reflecting the history of the development of the Yi society. Some examples of Yi characters are shown in Fig. 6. In this part, the main content of the handwritten Yi character recognition will be detailed, which is divided into two parts: the construction of the dataset and the experiment.

1) THE CONSTRUCTION OF THE DATASET

There are several steps in the construction of the dataset: First, crop images of Yi characters from scanned documents of Yi classics. Second, these pictures are then binarized and normalized. Third, the clustering algorithm is implemented using these image slices as the input. Finally, class labels are added to the clustering results. Fig. 7 shows a flow chart of the construction of the dataset. The Yi characters used in this paper are cut from scanned documents of Yi classics, and we apply normalization and binarization to these image slices. After these steps, the images we obtained are binary images of size 100 × 100.

2) IMPROVED DPC ON YI CHARACTER IMAGE DATASET

To evaluate the experimental results of the clustering model and other clustering models, in Table 3, we present a statistical description of the clustering accuracy of K-means, DBSCAN, DPC, and the improved DPC. The difference between the positive and negative accuracies in the table is the deviation from the case where the threshold is manually selected multiple times. The density peak clustering method based on information entropy improvement is not based on human intervention, so the accuracy deviation is relatively small.

V. CONCLUSION

In the era of the IoT, IoT devices by their nature collect or generate large amounts of data from the environment, which cannot be processed manually. Aiming at the problem that ancient book image recognition, retrieval and labeling require labeled datasets, we propose an unsupervised clustering of the ancient book images of characters based on improved density peak clustering. In density peak clustering, the similarity of class images is determined by the metric function of design and research, and the extension of class members is carried out. After each distance threshold enumeration traversal, the change in state of the information entropy is obtained. By analyzing the threshold number of suspicious classes, the prediction of the set classification is realized, and the clustering process of images is completed. The experimental results show that the improved method can effectively implement the clustering of images, and the obtained labeled images can then be used for further semantic analysis.

REFERENCES

[1] H. Gao, W. Huang, X. Yang, Y. Duan, and Y. Yin, "Toward service selection for workflow reconfiguration: An interface-based computing solution," *Future Gener. Comput. Syst.*, vol. 87, pp. 298–311, Oct. 2018. doi: 10.1016/j.future.2018.04.064.

[2] Y. Yin, Y. Xu, W. Xu, M. Gao, L. Yu, and Y. Pei, "Collaborative service selection via ensemble learning in mixed mobile network environments," *Entropy*, vol. 19, no. 7, p. 358, 2017. doi: 10.3390/e19070358.

[3] S. Li, J. Wen, F. Luo, T. Cheng, and Q. Xiong, "A location and reputation aware matrix factorization approach for personalized quality of service prediction," in *Proc. IEEE Int. Conf. Web Services (ICWS)*, Honolulu, HI, USA, Jun. 2017, pp. 652–659. doi: 10.1109/ICWS.2017.78.

[4] P. He, J. Zhu, Z. Zheng, J. Xu, and M. R. Lyu, "Location-based hierarchical matrix factorization for Web service recommendation," in *Proc. IEEE Int. Conf. Web Services (ICWS)*, Anchorage, AK, USA, Jun./Jul. 2014, pp. 297–304. doi: 10.1109/ICWS.2014.51.

[5] Y. Yin, L. Chen, Y. Xu, and J. Wan, "Location-aware service recommendation with enhanced probabilistic matrix factorization," *IEEE Access*, vol. 6, pp. 62815–62825, 2018. doi: 10.1109/ACCESS.2018.2877137.

[6] H. Gao, H. Miao, L. Liu, J. Kai, and K. Zhao, "Automated quantitative verification for service-based system design: A visualization transform tool perspective," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 28, no. 10, pp. 1369–1397, 2018. doi: 10.1142/S0218194018500390.

[7] H. Gao, K. Zhang, J. Yang, F. Wu, and H. Liu, "Applying improved particle swarm optimization for dynamic service composition focusing on quality of service evaluations under hybrid networks," *Int. J. Distrib. Sensor Netw.*, vol. 14, no. 2, pp. 1–14, 2018. doi: 10.1177/1550147718761583.

[8] P. Berkhin, J. Kogan, C. Nicholas, and M. Teboulle, "A survey of clustering data mining techniques," in *Grouping Multidimensional Data: Recent Advances in Clustering*. Berlin, Germany: Springer, 2006, pp. 25–71. doi: 10.1007/3-540-28349-8_2.

[9] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010. doi: 10.1016/j.patrec.2009.09.011.

[10] S. S. Singh and N. Chauhan, "K-means v/s k-medoids: A comparative study," in *Proc. Nat. Conf. Recent Trends Eng. Technol.*, vol. 13, pp. 13–14, 2011.

[11] J. Han, M. Kamber, and J. P. Professor, *Data Mining: Concepts and Techniques*, 3rd ed. San Mateo, CA, USA: Morgan Kaufmann, 2011. [Online]. Available: http://hanj.cs.illinois.edu/bk3/

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining (KDD)*, Portland, OR, USA, 1996, pp. 226–231. [Online]. Available: http://www.aaai.org/Library/KDD/1996/kdd96-037.php

[13] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, Philadelphia, PA, USA, Jun. 1999, pp. 49–60. doi: 10.1145/304182.304187.

[14] A. Hinneburg and H. Gabriel, "DENCLUE 2.0: Fast clustering based on kernel density estimation," in *Proc. Adv. Intell. Data Anal. VII, 7th Int. Symp. Intell. Data Anal. (IDA)*, Ljubljana, Slovenia, Sep. 2007, pp. 70–80. doi: 10.1007/978-3-540-74825-0_7.

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

[16] S. Wang, D. Wang, C. Li, Y. Li, and G. Ding, "Clustering by fast search and find of density peaks with data field," *Chin. J. Electron.*, vol. 25, no. 3, pp. 397–402, 2016.

[17] S. Wang, D. Wang, C. Li, and Y. Li, "Comment on 'clustering by fast search and find of density peaks,'" Jan. 2015. [Online]. Available: https://arxiv.org/abs/1501.04267

[18] R. Mehmood, S. El-Ashram, R. Bie, H. Dawood, and A. Kos, "Clustering by fast search and merge of local density peaks for gene expression microarray data," *Sci. Rep.*, vol. 7, Apr. 2017, Art. no. 45602.

[19] Y. Yin, L. Chen, Y. Xu, J. Wan, H. Zhang, and Z. Mai, "QoS prediction for service recommendation with deep feature learning in edge computing environment," *Mobile Netw. Appl.*, Apr. 2019. doi: 10.1007/s11036-019-01241-7.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015. doi: 10.1007/s11263-015-0816-y.

[21] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data?" *Perspect. Psychol. Sci.*, vol. 6, no. 1, pp. 3–5, Jan. 2011.

[22] A. Torralba, B. C. Russell, and J. Yuen, "LabelME: Online image annotation and applications," *Proc. IEEE*, vol. 98, no. 8, pp. 1467–1484, Aug. 2010. doi: 10.1109/JPROC.2010.2050290.

[23] G. Ciocca, P. Napoletano, and R. Schettini, "IAT—Image annotation tool: Manual," Feb. 2015, *arXiv:1502.05212*. [Online]. Available: https://arxiv.org/abs/1502.05212

[24] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognit.*, vol. 45, no. 1, pp. 346–362, 2012. doi: 10.1016/j.patcog.2011.05.013.

[25] P. Bi, "Handbook of linguistic annotation," *J. Quant. Linguistics*, vol. 25, no. 4, pp. 372–376, 2018. doi: 10.1080/09296174.2018.1424495.

[26] C. Cusano, G. Ciocca, and R. Schettini, "Image annotation using SVM," in *Proc. SPIE*, vol. 5304, pp. 330–339, Dec. 2003.

[27] E. Chang, K. Goh, G. Sychay, and G. Wu, "CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machines," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 26–38, Jan. 2003. doi: 10.1109/TCSVT.2002.808079.

[28] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1075–1088, Sep. 2003. doi: 10.1109/TPAMI.2003.1227984.

[29] C. Yang, M. Dong, and J. Hua, "Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2006, pp. 2057–2063. doi: 10.1109/CVPR.2006.250.

[30] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002. doi: 10.1109/34.1000236.

[31] C.-F. Tsai, "Bag-of-words representation in image annotation: A review," *ISRN Artif. Intell.*, vol. 2012, p. 19, 2012. [Online]. Available: http://downloads.hindawi.com/archive/2012/376804.pdf. doi: 10.5402/2012/ 376804.

[32] F. Perronnin, Y. Liu, and J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, Jun. 2010, pp. 3384–3391. doi: 10.1109/CVPR.2010.5540009.

[33] A. Hanbury, "A survey of methods for image annotation," *J. Vis. Lang. Comput.*, vol. 19, no. 5, pp. 617–627, 2008. doi: 10.1016/j.jvlc.2008.01.002.

[34] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3156–3164. doi: 10.1109/CVPR.2015.7298935.

[35] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Lille, France, Jun. 2015, pp. 2048–2057. [Online]. Available: http://jmlr.org/proceedings/papers/v37/xuc15.html

[36] L. Zhou, C. Xu, P. Koch, and J. J. Corso, "Watch what you just said: Image captioning with text-conditional attention," Jan. 2016, *arXiv:1606.04621*. [Online]. Available: https://arxiv.org/abs/1606.04621

[37] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2285–2294. doi: 10.1109/CVPR.2016.251.

[38] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 203–212. doi: 10.1109/CVPR.2016.29.

[39] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *Proc. 1st Int. Workshop Multimedia Intell. Storage Retr. Manage.*, 1999, pp. 1–9.

[40] V. N. Murthy, S. Maji, and R. Manmatha, "Automatic image annotation using deep learning representations," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Shanghai, China, Jun. 2015, pp. 603–606. doi: 10.1145/2671188.2749391.

[41] A. Bouzaieni and S. Tabbone, "Images annotation extension based on user feedback," in *Proc. Adv. Concepts Intell. Vis. Syst.-18th Int. Conf. (ACIVS)*, Antwerp, Belgium, Sep. 2017, pp. 418–430. doi: 10.1007/978-3-319-70353-4_36.

[42] D. Im and G. Park, "Linked tag: Image annotation using semantic relationships between image tags," *Multimedia Tools Appl.*, vol. 74, no. 7, pp. 2273–2287, 2015. doi: 10.1007/s11042-014-1855-z.

[43] Y. Yang, F. Wu, F. Nie, H. T. Shen, Y. Zhuang, and A. G. Hauptmann, "Web and personal image annotation by mining label correlation with relaxed visual graph embedding," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1339–1351, Mar. 2012. doi: 10.1109/TIP.2011.2169269.

[44] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, and A. D. Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval," *ACM Comput. Surv.*, vol. 49, no. 1, p. 14, 2016. doi: 10.1145/2906152.

[45] A. Tariq and H. Foroosh, "Learning semantics for image annotation," May 2017, *arXiv:1705.05102*. [Online]. Available: https://arxiv.org/abs/1705.05102

[46] R. A. Jarvis and E. A. Patrick, "Clustering using a similarity measure based on shared near neighbors," *IEEE Trans. Comput.*, vol. C-22, no. 11, pp. 1025–1034, Nov. 1973. doi: 10.1109/T-C.1973.223640.

YU WENG received the Ph.D. degree in computer science from the University of Science and Technology Beijing, China, in 2010. He is currently an Associate Professor of computer science with the Information Engineering Department, Minzu University of China. He has published more than 20 conference and journal papers. His current research interests include machine learning, cloud computing, and distributed computing.

NING ZHANG is currently pursuing the master's degree in information engineering from the Minzu University of China. His current research interests include machine learning and pattern recognition.

XIAOXIAN YANG received the Ph.D. degree in management science and engineering from Shanghai University, Shanghai, China, in 2017. She is currently an Assistant Professor with Shanghai Polytechnic University, China. Her research interests include business process management and formal method.