Classification of Hyperspectral image based on superpixel segmentation and DPC algorithm

Nian Chen
College of Big Data and Artificial Intelligence
ChiZhou University
ChiZhou, China
czxycn@czu.edu.cn

Hao Zhou
College of Computer and Information Science
Southwest University
ChongQing, China
Zhouhao19@email.swu.edu.cn

***Abstract*—In this paper, we propose an algorithm named SS_DPC for hyperspectral image classification. First, the image is segmented into hyperpixels according to spatial and spectral information, which are used as basic units for clustering instead of pixels. Computing the inner product of the local density and the minimum inter_cluster distance for each unit , Density peaks clustering (DPC) algorithm sorts products in descending order and selects the globally optimal solutions as cluster centers.The following conclusions have been verified through experiments:(1)Proper quantity of superpixel (K value) can improve the consistency between clustering results and actual values effectively.(2)Image segmentation can weaken the interference of abnormal data, so the ARI values of SS_DPC,SS_K_Means are higher than that of K_Means significantly.(3)SS_DPC algorithm is much better than other clustering algorithms in precision and robustness.**

Keywords-hyperspectral image; density peaks clustering; image segmentation;band selected; average spectrum

I. INTRODUCTION

Compared with color images,more spatial and spectral information is included in hyperspectral images(HSI).For instance,remote sensing imagery integrates geometrical characteristic and spectral feature which is distributed on hundreds of bands continuously,and we can use them for atmospheric analysis, ocean exploration, mineral resource estimation and so on.With more and more images being collected, lots of data will be stored and calculated,so the method that segments image into multiple superpixels instead of pixels can reduce the amount of data and improve speed of processing effectively.DPC algorithm which was proposed by Rodriguez and Laio[1] in 2014 can obtain the global optimal solutions without iteration and perform well in dealing with the data which has the characteristic of irregular geometric spatial distribution .

Superpixel is a collection of pixels with spatial proximity and spectral similarity in HSI.In recent years, lots of research work has been done to superpixel segmentation, and the technology has been widely applied in various fields.In 2015, Zhan[2] et al.used multiple_layer superpixel graph and loopy belief propagation to segment HSI. In 2016,uniformity based superpixel segmentation was proposed by Saranathan[3].As a new member of the density_based clustering methods, comparative advantages of DPC algorithm have been demonstrated in different applications,including Biomedical image processing[4], band selection for HSI[5] etc.

In this paper, a method named SS_DPC is proposed.First, the superpixel segmentation technology is used to extract the spectral features of HSI, and then DPC algorithm complete the object clustering with the superpixel as the basic unit.By comparing performance of SS_DPC,SS_K_Means,K_Means through experiments,we found facts as following:(1)superpixel segmentation can not only reduce the number of units to be classified, but also cut down the influence of noise effectively.(2)DPC algorithm is more adaptable to non_spherical spatial distribution,and it has significant advantages in precision and robustness.

II. THEORY

A Superpixel Segmentation

Traditional pixel_level HSI classification has some defects such as large computation, data redundancy and salt noise etc. Superpixel is an area composed of pixels with similar characteristics such as color, brightness and texture.Suppose that N pixels of a HSI $X = \{x_i\}_{i=1}^{N}$ are distributed on m bands, and $R = \{r_i\}_{i=1}^{K}$ is the collection of superpixels, where N/K is the average number of pixels included in a superpixel.

There are two types of superpixel segmentation methods: one is gradient_ descent_ based algorithm, such as watershed algorithm[6](WS), mean shift algorithm[7](MS), and simple linear iterative clustering algorithm[8](SLIC) etc,and the other is graph_ theory_ based algorithm,such as minimum spanning tree algorithm[9](MST),superpixel lattices[10] algorithm(SL),entropy rate based algorithm(ER) [11]etc.

SLIC is an efficient algorithm whose output has the characteristics of standardized, compact, and small computation. The spectral and geometric distance between pixel and superpixel should be considered simultaneously when segmentation is carried out.The objective function of SLIC can be expressed as :

$$\begin{aligned} F &= s\_dis + \frac{M}{S} \times g\_dis \\ s\_dis &= \| x_i - r_k \| = \sqrt{\sum_{j=1}^{B} (x_{ij} - r_{kj})} \\ g\_dis &= \sqrt{(x_u - r_u)^2 + (x_v - r_v)^2} \end{aligned} \qquad (1)$$

In the expression (1), we initialize the image to many grids with size of S×S,and parameter M controls the compactness between superpixels, which usually takes value in the interval [1,40]. Usually, we calculate the distance s_dis in the CIELAB color space instead of the sum of all bands. Therefore, SLIC can use a 5-dimensional vector to represent the spectral information and position

2473-3636/20/$31.00 ©2020 IEEE
DOI 10.1109/DCABES50732.2020.00044

information of a pixel on the plane, $x_j = [l_j, a_j, b_j, x_j, y_j]$ .Each superpixel is a basic cluster, and image I can be regarded as a set of clusters, $I = (c_1, c_2, \cdots c_k)$.

The purpose of segmentation is to segment the original image into K units with an appropriate size $S = \sqrt{N / K}$, and there should be clear boundaries between these units. The boundary accuracy is an important index to measure the segmentation algorithm.The value of parameter K has a great influence on the segmentation effect.When it's larger, the image may be over_segmented, that means an entire unit will be divided into smaller areas meaninglessly,and the time complexity of the algorithm will increase.On the contrary,if the value of K is too small, it may cause insufficient segmentation, and pixels with different attributes will be clustered together.

Setting K by manual experience value perhaps result in the decrease of boundary recognition accuracy or the increase of computational burden. In recent years, a lot of researches have been done on the determination of K , but this part will not be discussed in the paper.

B Density peaks clustering algorithm

Due to the narrow and overlapping spectra of adjacent HSI bands, high correlation and redundancy of data are brought. With the rising of the number of spectra, high dimensions lead computation complexity and storage consumption increasing continuously which is called "dimension disaster" . Therefore ,it's a effective way to reduce dimensions by selecting parts of the bands to reconstruct image. Ranking_based[12] method and density_based[13] method are used frequently to form subset of bands. Traditional methods such as maximum_variance PCA[14](MVPCA),Information divergence[(ID) and affinity propagation[15](AP) etc all can be used for band selected to achieve the goal of information dimension reduction.As one of density_based algorithms, in essence,DPC is looking for high_density areas split by low_density areas, and it can obtain higher accuracy and robustness without any labeled samples .

DPC algorithm is based on the following two assumptions: firstly, the density of the cluster center is higher than that of the surrounding points; secondly, the distance value between the cluster center and the higher density point is relatively large.The local density ρ_i of point x_i can be defined as expression (2):

$$\rho_i = \sum_{j=1, j\neq i}^{N} \chi(d_{ij} - d_c) \qquad (2)$$

$$\chi(x) = \begin{cases} 1 & x \le 0 \\ 0 & x > 0 \end{cases}$$

Count the number of points whose distance from point x_i is less than the cutoff distance d_c .In order to reduce the adverse effect of statistical error, gaussian kernel function can also be used to calculate local density.In expression (3), d_{ij} is Euclidean distance between two points, and d_c is the weight parameter which is used to avoid overlapping or sparse among clusters.

$$\rho_i = \sum_j e^{-(\frac{d_{ij}}{d_c})^2} \qquad (3)$$

Assuming that point x_i has the highest local density and x_j is the point farthest from x_i ,the inter_cluster distance is defined as $\delta_i = \max_j (d_{ij})$ .Different from the highest density point,the distance of inter_cluster for other points satisfies $\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij})$ .So ,we define δ_i as expression (4)

$$\delta_i = \begin{cases} \min\limits_{j:\rho_i < \rho_j} (dist(x_i, x_j)), \ \exists j \ s.t. \ \rho_i < \rho_j \\ \max\limits_j (dist(x_i, x_j)), \quad otherwise \end{cases} \qquad (4)$$

Sometimes, ρ and δ need to be normalized to maintain the same metrics.

$$\tilde{\rho}_i = (\rho_i - \rho_{\min}) / (\rho_{\max} - \rho_{\min})$$
$$\tilde{\delta}_i = (\delta_i - \delta_{\min}) / (\delta_{\max} - \delta_{\min}) \qquad (5)$$

After segmentation, the superpixels with similar spectral features are clustered together to realize object recognition in the image. Considering each superpixel as a point,for all $r_i \in R$ compute $\gamma_i = \tilde{\rho}_i \times \tilde{\delta}_i$ and sort them by descending order.p points (p<K)are selected as cluster centers from large to small according to γ . As shown in Fig.1 which is called decision graph,the cluster centers are usually some outliers.The normal points are close to the horizontal axis ρ , and the abnormal points are close to the vertical axis δ.The optimal points which are most suitable for cluster centers usually appear in the upper right corner of the graph, that is, they have both high local density and sufficient cluster spacing, for instance, point A in Fig1.The points in the upper left corner near the axis are probably labeled as noise ones.

Construct the distance membership matrix,and assign non_center point to a certain cluster who is closest to it and the local density is higher than it.

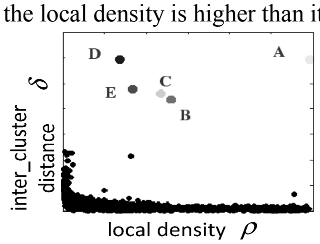

Fig. 1 Decision graph of DPC

III. ALGORITHM

SS_DPC algorithm is executed in three steps: first of all,SS algorithm segments HSI into multiple superpixels according to the spectral and geometric information.Then considering each superpixel as a basic unit, DPC is used to look for cluster centers.Finally, through comparing the distance between non_center point and every cluster center ,algorithm complete HSI object classification. In the paper,SLIC algorithm is carried out for superpixel segmentation,and each superpixel is represented by the average spectra of the pixels it contains.

DPC can works in both band selected and superpixel clustering.As one of unsupervised clustering algorithms,it has played good performance on data set whose geometry shape is non_spherical.In application of band selection, it's a effective way to use DPC to get representative bands from all bands, but the image recognition accuracy will be cut down due to the discarding of some bands.

Table1 Steps of DPC algorithm

Input: band set $B=(b_1,b_2,\cdots b_m)$,threshold α,β
Steps: 1.for each band $b_i \in B$ calculate ρ_i,δ_i according to expression (2) and (4); 2.if $(\rho_i<\alpha)$ and $(\delta_i>\beta)$, tag b_i as noise band 3.store $\gamma_i=\rho_i\times\delta_i$ to array T, sort elements in T by descending order $\gamma_1>\gamma_2>\cdots>\gamma_m$ 4.select $b_1\sim b_p$ corresponding to $\gamma_1\sim\gamma_p\,(p<m)$ takes $B'=\{b_1,b_2,\ldots b_p\}\subset B$ represent B
Output: subset $B'=\{b_1,b_2,\ldots b_p\}$

Table 1 shows the implementation steps of band selection with DPC algorithm, reducing the data dimension and eliminating the noise band on the premise that the main part of the information is not lost.

The execution steps of the SS_DPC algorithm are provided as Table 2:

Table 2 Steps of SS_DPC algorithm

Input:HSI pixels $\{x_i\}_{i=1}^{N}$, number of superpixels K , parameter d_c, p,q
pre_operation: for $\{x_i\}_{i=1}^{N}$,use DPC algorithm $x_i=(x_{i1},x_{i2},\cdots x_{im})\Rightarrow x_i=(x_{i1},x_{i2},\cdots x_{ip}),p<m$; step1: according to objective function $F=s\_dis+\frac{M}{S}\times g\_dis$ $\{x_i\}_{i=1}^{N}\overset{SLIC}{\Rightarrow}\{r_i\}_{i=1}^{K}\quad r_j=average\{x_j\}_{j=1}^{N/K}$; Step2: use DPC algorithm extract spectral cluster centers $\{C_j\mid j=1,2,\cdots q\}$, q is number of clusters; Step3: Compute distance $dis(r_i,C_j)$,construct membership matrix $U=[U_{ij}]_{K\times q}$;
Output：label for each r_i.

In Fig.2, HSI is firstly segmented into superpixels, and then clustering is carried out to complete the process of object recognition.

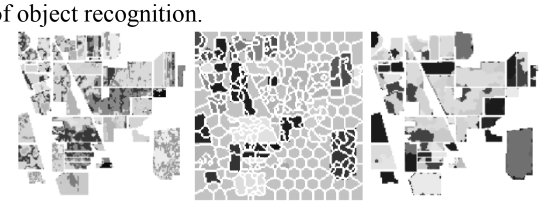

Fig. 2 The instance of SS_DPC algorithm

IV. Experiment and analysis

The experimental design of this paper includes two aspects: (1) Use DPC algorithm to extract subset of HSI bands, comparing the performance with MVPCA ,AP etc. (2) Use ARI indicator to analyze the execution effect of SS_DPC ,comparing it with K-Means ,SS_K_Means in precision and robustness.

The image size in Indian pines data set is 145×145 pixels which are distributed on 220 bands (range: 0.4um~2.5um).The spectral resolution of the image is 10nm and spatial resolution is 20m. According to the distribution of ground objects, Indian Pines remote sensing image sets 16 categories and assigns 10,366 labeled samples into them.We use DPC, PCA and AP algorithms to select bands from the Indian Pines samples respectively,and SVM model was used to classify the new samples which have been reduced the dimensions. Analyze the band selection performance of the above three algorithms by using OA and Kappa indicators.

Table 3 Performance comparison of different algorithms on band selection

algorithm	label of the band	OA	Kappa
DP	196,63,76,48,25	64.6%	0.62
	196,63,76,48,25,140,13,134	69.7%	0.68
MVPCA	66,27,144,192,31	62.1%	0.60
	66,27,144,192,31,176,75,16	64.8%	0.63
AP	12,35,75,117,183	64.4%	0.63
	12,35,42,75,97,117,132,183	70.2%	0.68
all _bands	-	82.4%	0.81

It can be seen from Table 3 that the overall accuracy of the band selection algorithm has been decreased comparing with the all_bands classification, but subset of the bands can still reflect the key physical features of objects in the image.The Kappa values of various algorithms are in the interval [0.60,0.65], which means that although the number of bands is very small(5 or 8), the classification results of algorithms can still maintain good consistency with the actual labels.With the increase of spectral dimensions, the OA and Kappa values of various algorithms are improved accordingly. The effect of DP and AP is similar, and both of them are better than MVPCA. The result indicates when the objects on the image are non_spherical spatial distribution, DPC algorithm can play its advantages in clustering.

Set the K value (number of hyperpixels) from 50 to 800, and compare the performance of SLIC algorithm under different segmentation granularity.ARI was used to measure the consistency between unsupervised clustering results and true situations.

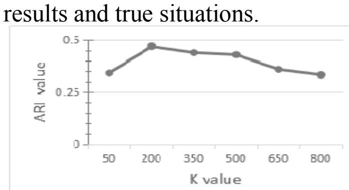

Fig. 3 Influence of K value in SLIC algorithm

In Fig.3, improper setting of K will result in value of ARI reduction. In the case of small K (less than 200), multiple clusters may overlap each other on the same superpixel. The shortage of basic units will lead to low boundary accuracy of superpixel segmentation .Similarly, large K will cause the image to be over segmented, and the computational complexity and error risk of clustering will increase.

Finally, K_Means ,SS_K_Means and SS_DPC algorithms are compared comprehensively. Only K_Means algorithm does not perform superpixel segmentation, but clustering pixels directly.In the experiment,we reconstruct Indian Pines image with 10 bands, and set K=200.

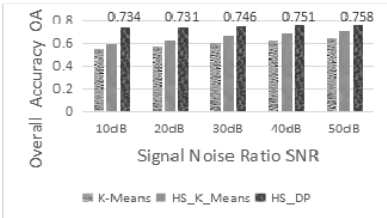

Fig. 4 Comparison of algorithm performance under different SNR environments

In Fig.4,SS_ DPC has the highest overall accuracy among the three algorithms ,because it extracts image features based on superpixels and scans the optimal clusters in the global scope. The influence to SS_DPC with different SNR is very small, which shows that the algorithm has strong ability to resist data volatility and noise disturbance, and has good robustness.Superpixel segmentation can effectively reduce the noise interference in hyperspectral image, but it also brings extra computational burden. So, K_ means operates on pixels directly, its accuracy is the lowest of the three algorithms, while its time consumption is the least.

V. CONCLUSION

Hyperspectral image combine the geometric and spectral characteristics of pixels, which can be used for automatic object recognition.In this paper, firstly, we segment the image to K hyperpixels,and K is a value that need to be predetermined. The pixels within the hyperpixel have the characteristics of adjacent space and similar spectrum,and the average spectrum of pixels represents the spectrum of the superpixel. Then the density peaks clustering algorithm is used to clustering the superpixels to separate the objects from the environment.DPC algorithm performs well to non_spherical spatial distribution data,and it has strong ability of outliers detection. For each point,by sorting the product of local density and minimum inter_cluster distance,the optimal cluster centers are found.In this paper, DPC algorithm is used in band selection and superpixels clustering respectively, both of which aim at finding globally optimal cluster centers.

SS_DPC algorithm adds some time consumption for image segmentation,but it is more suitable for irregular data distribution geometrical shape, with the characteristics of high recognition accuracy and strong robustness.

REFERENCES

[1]Rodriguez A, Laio ,"Clustering by fast search and find of density peaks,"Science,vol.344,Jun.2014,pp.1492-496,doi:10.1126/science.1242072.

[2]Zhan T,Xu Y,Sun L,"Hyperspectral image classification using multilayer superpixel graph and loopy belief propagation," Geoscience and Remote Sensing Symposium,Jul.2015,doi:10.1109/IGARSS.2015.7326112.

[3]Arun M,Parente M,"Uniformity based superpixel segmentation of hyperspectral images,"IEEE Transactions on Geoscience and Remote Sensing ,vol.54, Mar.2016,pp.1419-1430, doi:10.1109/tgrs.2015.2480863.

[4]Chen P,Fan X,Liu R,"Fiber segmentation using a density_peaks clustering algorithm.Biomedical Imaging,"Proc.IEEE 12th International Symposium on,IEEE Press,Apr.2015,pp.633-637, doi:10.1109/ISBI.2015.7163953.

[5]Sun K,Geng X,Ji L,"Exemplar component analysis:A fast band selection method for hyperspectral imagery,"IEEE Transactions on Geoscience and Remote Sensing,vol.12,May.2015,pp.998-1002,doi:10.1109/lgrs.2014.2372071.

[6]L.Vincent, P.Soille, "Watersheds in digital spaces:an efficient algorithm based on immersion simulations,"Proc.IEEE Transactions on Pattern Analysis and Machine Intelligence,vol. 13,Jun.1991,pp.583-598,doi:10.1109/34.87344.

[7]D.Comaniciu, P.Meer,"Mean shift: a robust approach toward feature space analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence,vol. 24,May.2002,pp.603-619,doi:10.1109/34.1000236.

[8]R.Achanta, A.Shaji, K.Smith,"SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," IEEE Transactions on Pattern Analysis and Machine Intelligence,vol. 34 ,Nov.2012,pp.2274-2282,doi:10.1109/TPAMI.2012.120.

[9]C.Cevahir,A.A.Aydin,"Efficient graph-based image segmentation via speeded-up turbo pixels,"Proc.IEEE International Conference on Image Processing, IEEE Press,Sept. 2010,pp.3013-3016,doi:10.1109/ICIP.2010.5653963.

[10]A.P.Moore,S.J.D.Prince,J.Warrell,"Superpixel lattices,"Proc. IEEE Conference on Computer Vision and Pattern Recognition,IEEE Press,Jun.2008,pp.1-8,doi:10.1109/CVPR.2008.4587471.

[11]M.Liu, O.Tuzel, S.Ramalingam, "Entropy-based Clustering : Cluster Analysis via Maximizing a Submodular Function Subject to a Matroid Constraint," IEEE Transactions on Pattern Analysis and Machine Intelligence,vol. 36,Jan.2014,pp.99-112,doi:10.1109/TPAMI.2013.107.

[12]C.I.Chang, S.Wang, "Constrained band selection for hyperspectral imagery," IEEE Transactions on Geoscience and Remote Sensing, vol. 44,Jun.2006,pp.1575-1585,doi:10.1109/tgrs.2006.864389.

[13]A.Martinez,F.Pla, J.M.Sotoca,"Clustering-based hyperspectral band selection using information measure," IEEE Transactions on Geoscience and Remote Sensing,vol. 45,Dec.2007,pp.4158-4171,doi:10.1109/TGRS.2007.904951.

[14]C.I.Chang, Q.Du, T.L.Sun,"A joint band prioritization and band-deccorelation approach to band selection for hyperspectral image classification,"IEEE Transactions on Geoscience and Remote Sensing, vol.37 ,Nov.1999,pp.2631-2641.

[15]J.F.Frey, D.Dueck,"Clustering by passing messages between data points,"Science ,Feb.2007,pp.972-976,doi:10.1126/science 1136800.