MINERAL MAPPING WITH HYPERSPECTRAL IMAGE BASED ON AN IMPROVED K-MEANS CLUSTERING ALGORITHM

Zhongliang Ren[1], *Lin Sun*[1*], *Qiuping Zhai*[2], *XiRong Liu*[1]

[1]Geomatics College, Shandong University of Science and Technology, Qingdao 266590, China

[2]Shandong provincial key laboratory of water and soil conservation and environmental protection, college of resources and environment, Linyi University, Linyi 276000, China

ABSTRACT

Mineral mapping with hyperspectral images has been demonstrated as an effective way for land resources survey. K-means, as a typical clustering algorithm, is commonly used to process the object identification of hyperspectral images. However, due to the influence of mixed pixel, the matching of data points and cluster centers of the traditional k-means clustering algorithm is very difficult. Therefore, this paper proposes an improved k-means clustering algorithm to identify the mineral types from the AVIRIS hyperspectral image of Cuprite mining area. This algorithm uses three methods to select the initial cluster centers and spectral information divergence instead of Euclidean distance for better measuring the similarity. Finally, by matching the clustering results with the mineral distribution map of this region and USGS mineral spectral library, it was found that the improved k-means clustering algorithm can get better clustering results and higher mineral mapping accuracy than the traditional algorithm.

Index Terms**—**k-means, mineral mapping, spectral information divergence, hyperspectral images

1. INTRODUCTION

Mineral is an important part of land resources. Traditional mineral mapping is mainly completed through ground survey, which requires extensive labor and time. With the development of remote sensing (RS) technology, mineral mapping based on hyperspectral images has been widely used [1]. Compared with the multispectral image, hyperspectral image which has wider spectral range and higher spectral resolution can realize rapid identification of mineral [2].

Unsupervised classification which is also known as clustering is a common method for the object identification of hyperspectral images. Clustering is a typical data mining method that groups similar objects into the same group without prior knowledge. K-means algorithm is a classical clustering algorithm [3-4] and one of the top ten classical data mining algorithms [5]. The traditional k-means clustering algorithm (DKM) is simple, fast and easy to implement. So it has been widely used in image classification, image segmentation, pattern recognition and other fields. However, this algorithm is sensitive to the initial cluster centers. Besides, the similarity measurement function also affects the results. The traditional k-means algorithm usually uses the Euclidean distance (ED) to measure the similarity of data points and cluster centers. The Euclidean distance does not eliminate the effect of dimensionality and is sensitive to the size difference of reflectivity. Moreover, limited by the spatial resolution of the sensors, the interpretation of RS is always influenced by mixed pixels. The phenomena of "same object with different spectrum" and "different objects with same spectrum" are common in RS images, which pose challenge to correct interpretation. Therefore, the clustering accuracy of hyperspectral images by DKM is generally poor.

Compared with the Euclidean distance, spectral information divergence (SID) is a probabilistic method that allows the variations of pixel values [6]. In addition, according to the three-sigma criteria, data points are mainly distributed in the range of (u-σ,u+σ). Therefore, this paper proposes an improved k-means algorithm (SKM) which uses SID to measure the similarity and three methods to initialize the cluster centers. Moreover, for a more comprehensive analysis of SKM clustering results, the improved k-means algorithms including SCM and MKM which respectively use spectral correlation distance (SCD) and mahalanobis distance（MD）instead of ED are performed.

2. METHODOLOGY

The basic principle of k-means clustering algorithm is to determine K cluster centers from a given dataset according to some criterions and iteratively calculate the cluster centers until the termination condition is met. The DKM usually uses the minimum intra-class variance as the clustering standard [7]. The workflow of k-means clustering algorithm is as follows:

978-1-5386-9154-0/19/$31.00 ©2019 IEEE

1) Initialize the cluster centers: randomly select K cluster centers from the dataset;
2) Clustering division: calculate the distance between data points and cluster centers. Then divide each data point into the nearest cluster center;
3) Update the cluster centers: calculate the average value of data points as the new cluster centers;
4) Iteration: when the specified iteration number is reached or the cluster centers are no longer changed, the iteration stops; Otherwise, go back to step 2 and continue the iteration.

Although the traditional k-means algorithm is simple and easy to implement, it is very sensitive to the initial cluster centers [8]. Moreover, the Euclidean distance only determines the similarity from the size of reflectivity value, which will lead to misclassification when the waveform is similar but the reflectivity size varies greatly. By measuring the divergence between data points and cluster centers, SID is not sensitive to the difference of reflectivity. Therefore, this paper uses SID instead of the Euclidean distance for better measuring the similarity. The formula of SID is shown as below [9].

$$SID(X,Y)=\sum_{i=1}^{n}(p(X_i)-q(Y_i))\log(\frac{p(X_i)}{q(Y_i)})$$

In this formula, X and Y are two spectral vectors. The value of SID is between 0 and 1. The smaller the SID value the more likely the data points and cluster centers are similar.

According to the distribution rule of random variables, the probability that the data value is distributed within the range of (u-σ,u+σ) is 0.6827. By selecting the initial cluster centers within this range, not only the number of iterations will be reduced, but outliers can be avoided as the cluster centers. The improved algorithm will be more efficient and stable than DKM.

The k-means clustering results only reflects the distribution of minerals and cannot accurately give the mineral types of the cluster centers. In order to get final mineral mapping results, this paper uses the spectral angle matching (SAM) to measure the similarity between the spectral curves of cluster centers and USGS mineral spectral library. SAM, as a classical matching method, measures the similarity of two spectral vectors by calculating the angle between them.

3. EXPERIMENT

3.1 RESEARCH AREA OVERVIEW

The research area of this paper is Cuprite mining area (figure 1) in Nevada, USA. Due to the high degree of mineral exposure and the low vegetation coverage in this area, many scientists use the hyperspectral images of this area to identify the types of minerals and achieve good results [10]. As can be seen from figure 1, which is taken as the reference mineral distribution map and has been registered with the hyperspectral image with an error of less than 1 pixel, Cuprite mainly includes alunite, kaolinite, montmorillonite, calcite, muscovite, chlorite and other minerals.

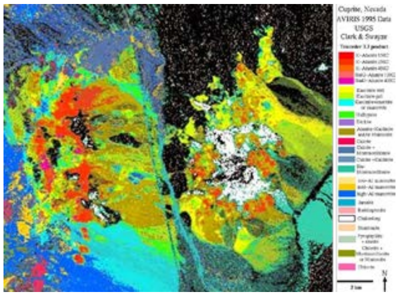


Figure.1. Mineral distribution map of Cuprite mining area

AVIRIS hyperspectral image of this area was selected as experimental data in this paper. AVIRIS image whose spectral resolution is 10nm and spatial resolution is 20m has 224 consecutive narrow bands and covers the wavelength range of 0.4um~2.5um. Due to its high spectral and spatial resolution, this data has been widely used in mineral mapping studies [11]. In order to identify various mineral types, USGS mineral spectral library is selected for spectral matching after clustering. It covers the spectrum range of 0.2-3.0um and includes the spectral curves of 481 typical minerals.

3.2 DATA PREPROCESSING

In order to eliminate the influence of atmospheric scattering and absorption, this paper uses FLAASH module of ENVI to conduct atmospheric correction of AVIRIS hyperspectral image. The spectral curves of the same pixel before and after atmospheric correction are shown in figure 2.

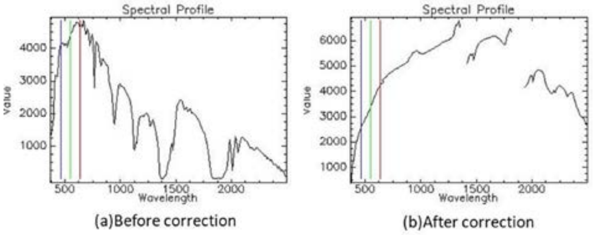


Figure.2. Spectral curves of the same pixel before and after atmospheric correction

It can be seen that AVIRIS data after atmospheric correction has two obvious water vapor absorption bands near 1450nm and 1950nm, which cannot be used. The spectral absorption bands of various minerals in Cuprite mining area are mainly concentrated in the short-wave infrared range of 1300nm~2500nm. Therefore, fifty short-wave infrared bands within the range of 2.0~2.5um were selected for mineral mapping in this paper.

Due to the differences in wavelength range and spectral resolution between AVIRIS hyperspectral data and USGS mineral spectral library, spectral resampling of the spectral library is required before matching. The spectral curves of Alunite in the spectral library before and after spectral resampling are shown in figure 3. The red line and the blue circles are the spectral curves of alunite before and after resampling respectively.

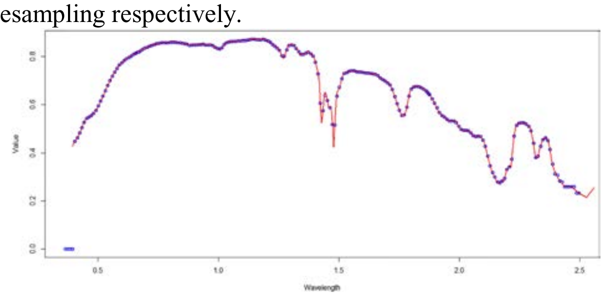
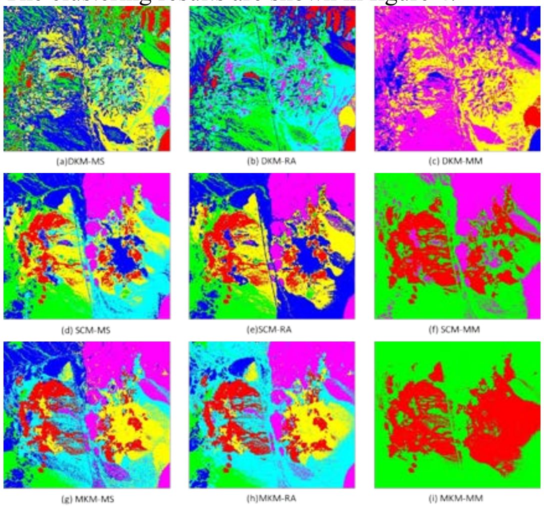


Figure.3. Spectral curves of Alunite before and after spectral resampling

3.3 Analysis of results

In order to analyze the influence of the initial cluster centers on the clustering results, three initialization methods which include random selection method (RA), max-min selection method (MM) and mean-sigma selection method (MS) were adopted in this paper. The max-min selection refers to the selection of initial cluster centers between the minimum and maximum values of the data's respective dimensions. And in order to weaken the influence of the phenomena of "same object with different spectrum" and "different objects with same spectrum" on DKM, this paper uses SID instead of ED for clustering analysis. Meanwhile, the comparison with other distance classifiers including SCD and MD are also performed. According to the mineral distribution map in figure 1, there are mainly 6 minerals in the Cuprite mining area. Therefore, the number of cluster centers K is set as 6. The clustering results are shown in figure 4.



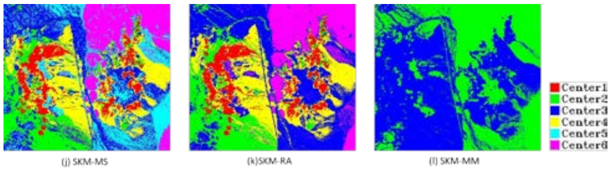


Figure.4. K-means clustering results

Each color in figure 4 represents a cluster center. Combined with figure 1, it can be seen that the mineral clustering results of DKM are disordered, while the results of the other three clustering algorithms can well reflect the distribution of various minerals. This shows that compared with the traditional k-means, the improved k-means based on SID, SCD and MD can get better clustering results. And in three initialization methods, MS and RA have better clustering results because they can generally reflect five or six minerals' distribution, while MM method can only reflect two minerals distribution.

The spectral curves of cluster centers were matched with USGS mineral spectral library by using SAM to get the mineral mapping results (figure 5).

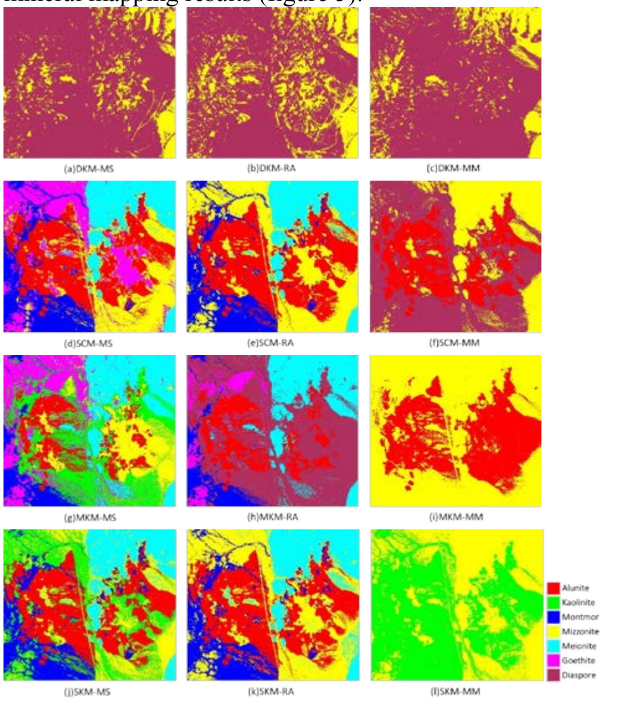


Figure.5. K-means mineral mapping results

In order to quantitatively analyze the clustering results of different k-means algorithms, random sample points were selected from the image and mineral distribution map, and the recognition accuracies were calculated (table 1).

Table 1 Recognition accuracies of different algorithms

Algorithms		AC
DKM	MS	0.00
	RA	0.00
	MM	0.00
SCM	MS	0.47
	RA	0.48

	MM	0.24
MKM	MS	0.41
	RA	0.20
	MM	0.23
SKM	MS	0.58
	RA	0.46
	MM	0.36

From figure 5 and table 1, it can be seen that no matter which initialization method is adopted, six minerals are misidentified as Mizzonite and Diaspore by DKM. For the other three algorithms, SKM has higher recognition accuracy. And MS generally has better clustering results in three initialization methods. Among all the clustering algorithms, SKM-MS which can get three minerals including Alunite, Kaolinite and Montmorillonite has the highest recognition accuracy which is 0.58. However, no method can identify all minerals because the spectral curves of minerals are very similar and RS images generally have many mixed pixels for its limited spatial resolution.

Meanwhile, by comparing the clustering results and mineral mapping results of SKM-MS and SKM-RA respectively, it is found that the number of minerals obtained by these two methods is one less than the number of the cluster centers. This is because the first cluster center (Center1) and the fourth cluster center (Center4) are matched into Alunite (figure 6).

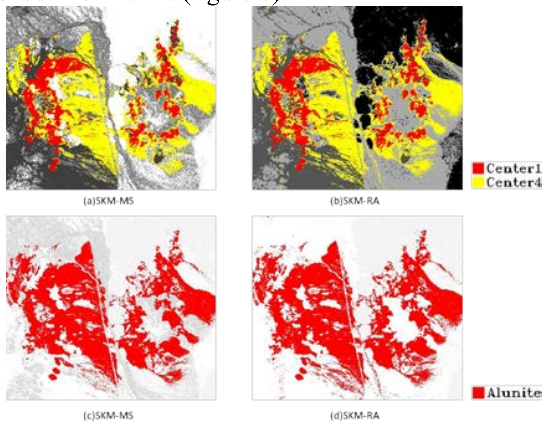


Figure.6. Alunite mapping results

As can be seen from the mineral distribution map in figure 1, the fourth cluster center is actually a mixed pixel which includes Alunite, Kaolinite and Muscovite. However, only Alunite has been identified by matching with the spectral library. This is because the USGS mineral spectral library contains pure spectrum. Therefore, each pixel in the hyperspectral image can only be divided into a specific mineral by matching with the spectral library.

4. CONCLUSION

The clustering results of traditional k-means are generally poor because the ED does not take into account the order of magnitude of each variable. Therefore, this paper uses SID to replace the ED for clustering analysis. And the comparison with SCD and MD are performed. Meanwhile, three methods were used to analyze the effect of initial cluster centers on clustering results. The results show compared with the traditional k-means clustering algorithm, the improved k-means clustering algorithms based on SID, SCD, and MD have better clustering results. Meanwhile, SKM has higher recognition accuracy than SCM and MKM. Besides, in three initialization methods, the clustering result of MS is better than the others. However, due to the influence of mixed pixels, some minerals are not identified. Therefore, in future studies, we can improve the accuracy of mineral mapping by end-element extraction and abundance inversion from the perspective of sub-pixel.

5. REFERENCES

[1] Vaughan, R.G., Calvin, W.M., & Taranik, J.V.. SEBASS hyperspectral thermal infrared data: Surface emissivity measurement and mineral mapping. Remote Sensing of Environment, 2003, 85(1): 48-63.

[2] Van der Meer, F., & Bakker, W.. Cross Correlation Spectral Matching: application to surface mineralogical mapping using AVIRIS data from Cuprite, Nevada. Remote Sensing of Environment, 1997, 61: 371-383.

[3] T. Kanungo, D.M Mount. A local search approximation algorithm for k-means clustering [J]. Computational Geometry，2004，28（2/3）：89-112.

[4] C. Elkan, Using the Triangle Inequality to Accelerate k-Means [C]. Proceedings of the 2nd International Conference on Machine Learning(ICML-2003). Menlo Park：AAAI Press, 2003：147-153.

[5] Wu X , Kumar V , Quinlan J R , et al. Top 10 algorithms in data mining[J]. Knowledge and Information Systems, 2008, 14(1):1-37.

[6] Amer R, Al Mezayen A, Hasanein M. ASTER spectral analysis for alteration minerals associated with gold mineralization. Ore Geology Reviews, 2015, 75(2016): 239-251.

[7] F.T. Grigorios, C.L. Aristidis, The MinMax k-Means clustering algorithm. Pattern Recognition, 2014, 47: 2505-2516.

[8] F.T. Grigorios, C.L. Aristidis, The global kernel k-means algorithm for clustering in feature space. IEEE TRANSACTIONS ON NEURAL NETWORKS, 2009, 20(7): 1181-1194.

[9] Qin J, Burks T F, Ritenour M A, et al. Detection of citrus canker using hyperspectral reflectance imaging with spectral information divergence. Journal of Food Engineering, 2009, 93(2):183-191.

[10] Chen X, Warner T A, Campagna D J. Integrating visible, near-infrared and short-wave infrared hyperspectral and multispectral thermal imagery for geological mapping at Cuprite, Nevada[J]. Remote Sensing of Environment, 2007, 110(3):344-356.

[11] Simon J. Hook, et al. An evaluation of short-wave-infrared (SWIR) data from the AVIRIS and GEOSCAN instruments for mineralogical mapping at Cuprite, Nevada. GEOPHYSICS, 1991, 56(9): 1432-1440.