Sensors & Transducers, Vol. 177, Issue 8, August 2014, pp. 128-135

Sensors & Transducers

http://www.sensorsportal.com

Cluster Validity Classification Approaches Based on Geometric Probability and Application in the Classification of Remotely Sensed Images

[1] LI Jian-Wei, [2] LI Xiao-Wen, [3] MAO Zheng-Yuan, [4] KONG Xiang-Zeng
[1] College of physics and information engineering, Fuzhou University, Fuzhou 350108 China
[2] College of mathematics and computer science, Longyan University, Longyan 364012 China
[3] Key Laboratory of Spatial Data Mining and Information Sharing of Ministry of Education, Spatial Information Research Center, Fuzhou University, Fuzhou 350002 China
[4] School of Computing and Mathematics, University of Ulster at Jordanstown, Newtownabbey, Northern Ireland, UK, BT37 0QB
[1] Tel.: +86-13950285662
[1] E-mail: lwticq@163.com

Received: 20 May 2014 /Accepted: 31 July 2014 /Published: 31 August 2014

Abstract: On the basis of the cluster validity function based on geometric probability in literature [1, 2], propose a cluster analysis method based on geometric probability to process large amount of data in rectangular area. The basic idea is top-down stepwise refinement, firstly categories then subcategories. On all clustering levels, use the cluster validity function based on geometric probability firstly, determine clusters and the gathering direction, then determine the center of clustering and the border of clusters. Through TM remote sensing image classification examples, compare with the supervision and unsupervised classification in ERDAS and the cluster analysis method based on geometric probability in two-dimensional square which is proposed in literature 2. Results show that the proposed method can significantly improve the classification accuracy.

Keywords: Two-dimensional rectangular area, Cluster validity, Cluster analysis, Geometric probability, Remote sensing image classification.

1. Introduction

Clustering relies on the similarity between things as generic grading criteria, distinguish and classify things in accordance with certain requirements and laws without any priori knowledge about the classification. It is one of the important directions of data mining and pattern recognition and plays an extremely important role of identifying the internal structure of data [3-5]. It is widely used in biology, climate science, economics and remote sensing, with the aim to distinguish between different things and recognize the similarities between things [6-9]. Therefore, the study of cluster analysis has important significance. Based on the clustering validity function based on geometric probability proposed by literature [1], this paper proposes a cluster analysis method to process large data in rectangular area, and applied it in remote sensing images classification to test its effectiveness.

2. Clustering Validity Classification based on Geometric Probability

2.1. Clustering Steps

"The idea of clustering method based on geometric probability is first dividing category, then divide subcategories. It is a top-down refining process. Based on the first cluster level, the cluster process operates only once. Assume that in the first cluster level, the sample is divided into c classes, and in the second cluster level, we only need to cluster each in c subclass once, so the same clustering process was repeated c times. So the times of clustering on each level cluster are equal to the number of subclass obtained by the upper level. If found that the subclass should not be subdivided, the clustering of this subclass is terminated. If all subclasses on a clustering level can not be subdivided, clustering finish" [2]. Each clustering on every clustering level includes structure plot, class determination, gathering direction determination, gathering center determination, class boundaries determination and classify (as Fig. 1).

Constructing scatter is the process to map two-dimensional data set to a rectangular area on a two-dimensional plane. Actually we only need to use the two-dimensional attribute data as two-dimensional coordinate data, and represent intuitively by electronic display devices. Class judgment and gathering direction determination is judged by the graph which generates function $H(\theta,\Delta\theta)$ of the data set. The standard to judge whether data gathering has subclass depends on whether there's distinct peaks in function $H(\theta,\Delta\theta)$. Gather direction is extracted according to the peak of function $H(\theta,\Delta\theta)$. The process of defining gathering center, defining boundary of class and classifying is as followed. Scan the point set in the rectangle area on two-dimensional plane, calculate the density of points. When the density of points reaches the peak value, the node of two directions is gathering center. After merge duplicate records of gathering center, connect gathering center one by one, find the valley of the point's density, and connect local density valley point from this point, then become the boundary of class. Mark category of the points according to the boundary, then get the classification result.

2.2. Cluster Validity Function Based on Geometric Probability

In the clustering process based on geometric probability, we research on class identification, class level and classify suspending judgment, gathering center identification on each level.

Whether there is peak of the figure of function $H(\theta,\Delta\theta)$ is the basis of judging data gathering. The peak is more obvious, the possibility of dataset contains subclass is greater. Fig. 2 shows two scatters (one includes two gathering centers and the other has on gathering center) and two figures of their corresponding function H. The difference is obvious from both intuitive feeling and quantitative measure.

Number of clusters is equal to the number of cluster centers. Using the method described in § 2.1, we obtain the clustering center of each clustering level and have the number on of each clustering level as the basis for testing clustering results.

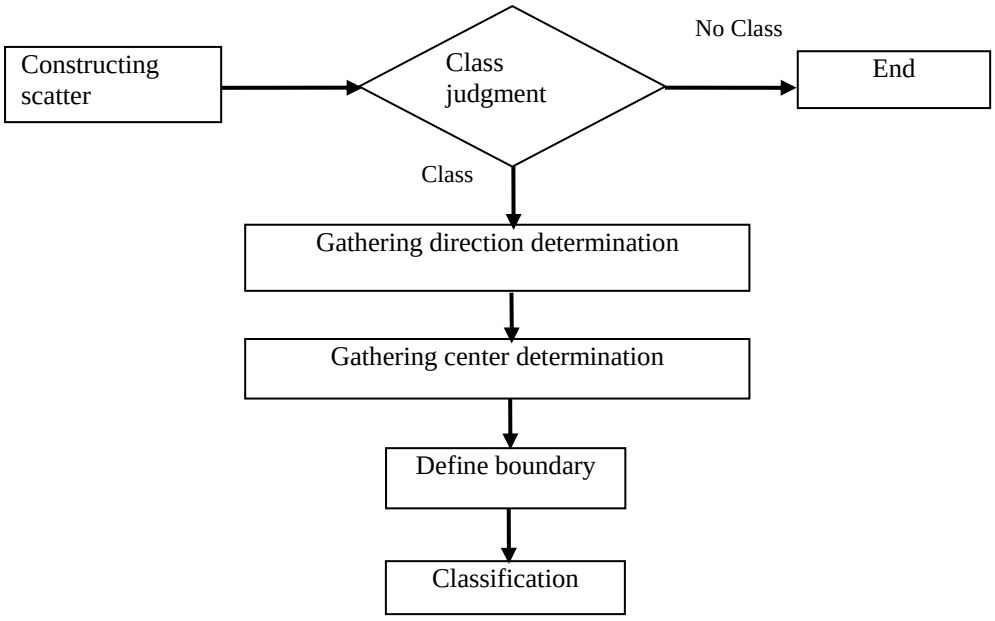
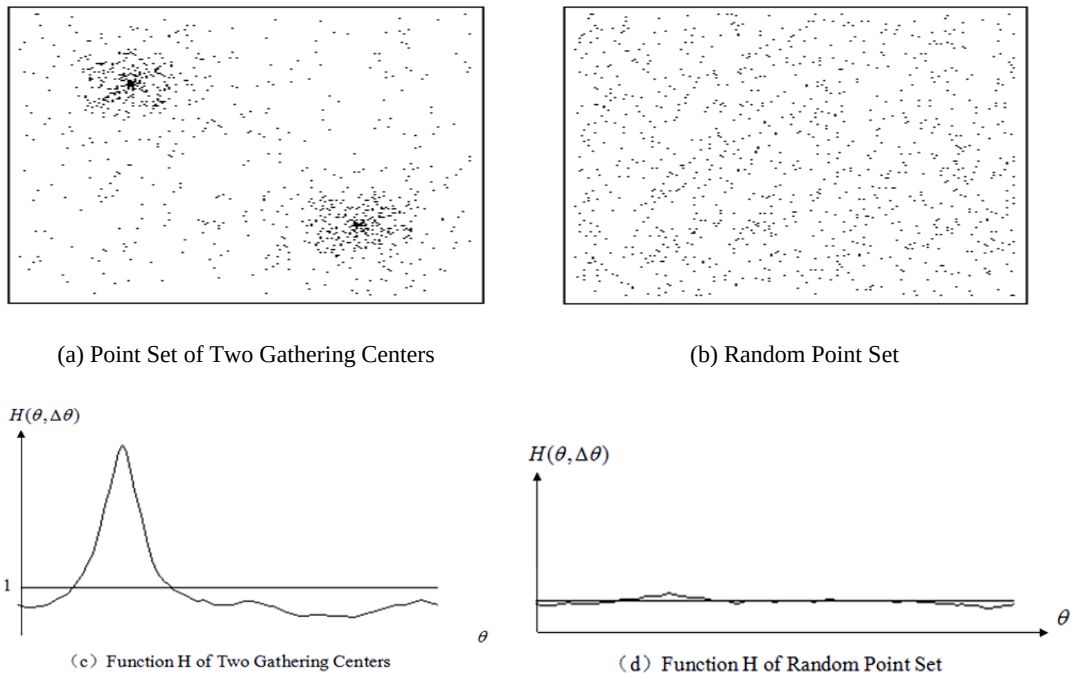


Fig. 1. Classification flowchart based on geometric probability.



(a) Point Set of Two Gathering Centers

(b) Random Point Set

(c) Function H of Two Gathering Centers

(d) Function H of Random Point Set

Fig. 2. Point set of two gathering centers and random point set and the figures of their function $H(\theta,\Delta\theta)$.

3.1. Clustering Process

Remote sensing image classification is one of the important topics of cluster analysis. Deviations of sensor, disturbance of weather changes and the complexity of terrain spatial distribution and seasonal change, so remote sensing images generally have the problem of same thing has different spectrums and different things have same spectrums. Remote sensing data belongs to large complex noisy sample. Using existing clustering algorithm directly, there are many errors in classification results. The accuracy is generally lower than supervised classification accuracy. In this section, we choose 1498 × 1281 pixels from Xiamen Jiulong River Estuary TM images of 2002 as experimental data source.

TM image data contains seven bands, is a seven-dimensional data set. We use part of the data in each clustering based on geometric probability. Data selection is highly flexible. In the subdivision process of different levels or different subclasses on same levels, we can choose two same bands or two different bands (or combination band of multiple bands). We can also choose one same band and one different band (or combination band of multiple bands). Which bands are chosen relates to the tasks on the particular classification level. In this section, we determine the data which clustering need on different clustering levels based on experience and spectrum feature.

3.1.1. Clustering on the First Level

The geographic objects which correspond with experimental data include water bodies, vegetation, residential areas, roads, farmland and unused wasteland. The first level of clustering choose data of the 4th and 5th band in remote sensing images of test area, where the gray value range of fourth band pixels are from 0 to 231, the gray value range of fifth band pixels are from 0 to 145, the fifth band is the x-axis, the fourth band is the y-axis, the scatter corresponding to the two-dimensional data set is in a rectangular area whose side lengths are a, b (a=145, b=231) (Show as Fig. 3).

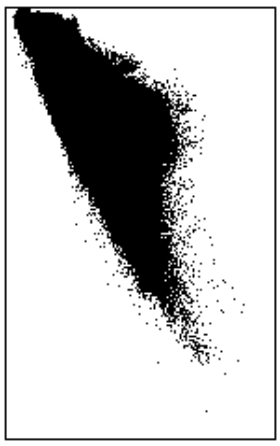


Fig. 3. Pixels Scatter.

According to the pixels scatter and the spatial clustering validity function H based on geometric probability, image of function H can be obtained (shown as Fig. 4), where the abscissa is the angular θ (degrees), the vertical coordinate value is value of H. It can be determined that there is one gathering

direction of the data set and the gathering direction is 0.57 (the angle with the y-axis positive). From the pixel bitmap shown as Fig. 3, we can find that gathering direction 0.57 radians is consistent with the gathering direction of actual data points.

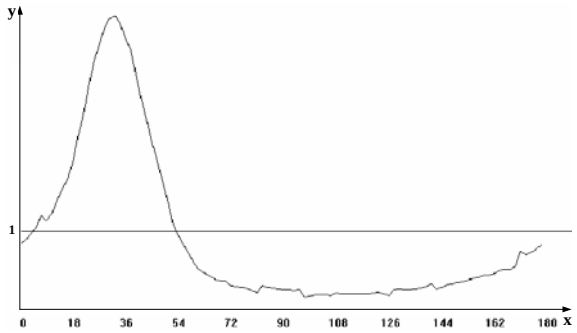


Fig. 4. Image of Function H.

Scan along the gathering direction and get the agglomeration cell density curve (Shown as Fig. 5). Then scan perpendicular to the gathering direction and get the curve (Shown as Fig. 6). There is only one peak in Fig.4 and Fig.5 means there has the highest density, two peaks of Fig. 6 illustrates there are two gathering center in the highest density parallel. The two gathering centers are (9, 6), (47,62) respectively.

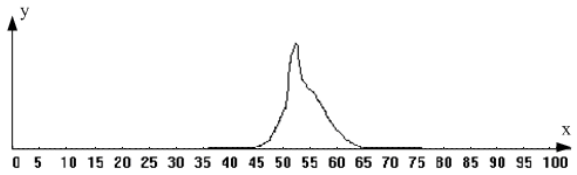


Fig. 5. Change of pixel density along the gathering direction.

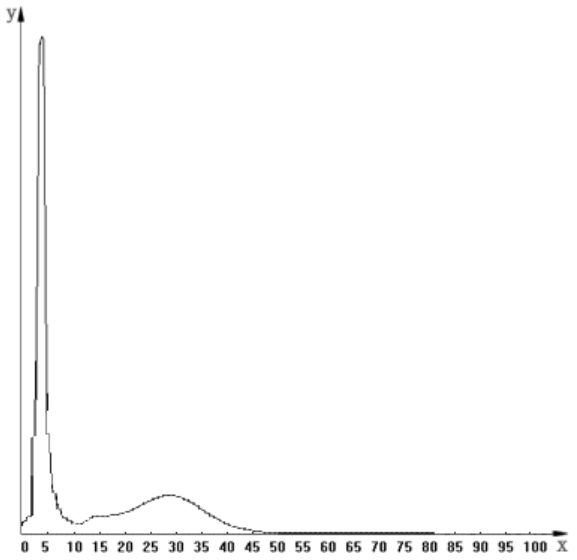


Fig. 6. Change of pixel density perpendicular to the gathering direction.

Connect two gathering center (9,6), (47,62) into a line, find the point on the line segment at the minimum density, set herein as a starting point, search for the valley point of local density on both sides along the direction perpendicular till reach the boundary of the feature space. At last connect each valley point of local density and get the dividing line (shown as Fig. 7).

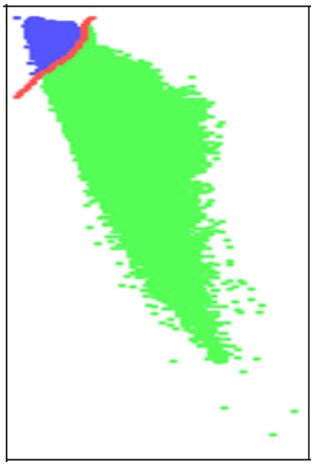


Fig. 7. Dividing line of classes.

The clustering results on the first level are displayed in remote sensing images, the upper left pixels are shown in blue, the lower right pixels are shown in green (shown as Fig. 7), their geographic boundaries generally consistent with the borders between land and sea, the black area is substantially water body (hereinafter referred to as A_1), the white area is substantially land (hereinafter referred to as A_2), shown as Fig. 8.

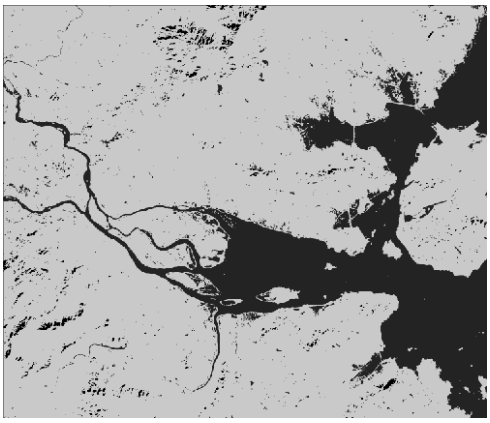


Fig. 8. Obtained class according to the dividing line.

3.1.2. Clustering on the Other Level

Continue to subdivide the two classes obtained on the fist level using the methods proposed in

Section 3.1.1. Use TM band 1 and band 2 as a data source on the 2nd level of Class A_1, the gathering direction is 1.14, gathering centers are (51, 20) and (71, 35), A_1 is divided into water body(hereinafter called B_1) and shadow; Use TM band 3 and band 4 as a data source on the 2nd level of Class A_1, the gathering direction is 0.36, gathering centers are (26, 10) and (38, 13), class B_1 is divided into clear water and turbid water; Use TM band 3 and the Synthetic band of band 3 and 4 as a data source on the 2nd level of Class A_2, the gathering direction is 2.8, the gathering centers are (19, 166) and (25, 182), A_2 is classified into non-vegetation (hereinafter called B_2) and vegetation (hereinafter called B_3).

Thereafter, subdivide class B_2 on the 3rd level using Band 4 and 5 of TM data, the gathering direction indicated by function H is 0.4, the gathering center are respectively (31, 27) and (47, 67), so we can divide B_2 into residential areas (hereinafter called C_1) and dry land (hereinafter called C_2). Finally, subdivide class B_3 on the 3rd level using Band 4 and 5 of TM data, the gathering direction indicated by function H is 0.44, the gathering center are respectively (36, 31) and (55, 68), and it is divided into dense vegetation and sparse vegetation.

3.2. Clustering Results and Comparative Analysis

Fig. 9 is a synthetic false color remote sensing images based on the TM data of Band 5, 4 and 3. After the clustering with our proposed algorithm, there are seven categories, including residential areas, shadows, close planting is, dilute vegetation, water, Cho and dry land (shown as Fig. 10).

Fig. 9. Original image.

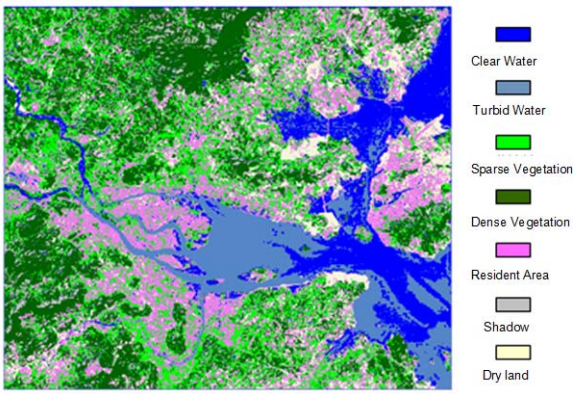


Fig. 10. Classification results.

According to the clustering results (shown as Fig. 11) of literature [2] and ERDAS, we conduct supervised classification and unsupervised classification to the same image separately. We pick up 250 pixels in the classification results to accuracy assessment, and compare with the clustering results of literature [2]. Upon examination, specific classification results of four clustering methods are shown as Table 1, Table 2, Table 3 and Table 4. The clustering methods include ERDAS supervised classification, ERDAS unsupervised classification, clustering method of literature [2] and clustering method of this paper. Bold figures in the table indicate the number of correct classification pixels for each class. Then calculate classification accuracy and the specific quantitative indicators using the figures in the four tables, results are shown in Table 5, Table 6, Table 7 and Table 8.

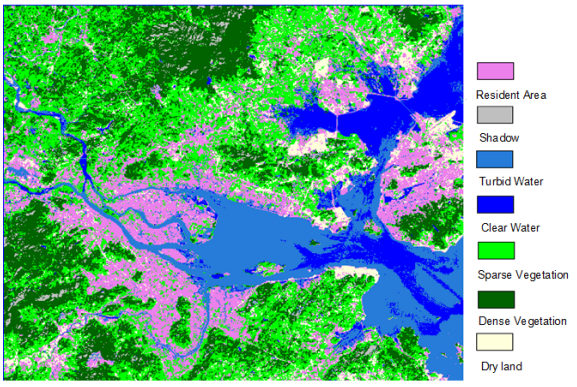


Fig. 11. Clustering results of literature.

Table 5, Table 6, Table 7 and Table 8 shows that the classification accuracy of the cluster validity method based on geometric probability is 85.20 %, while the classification accuracy of the method, supervised classification and unsupervised classification which are used in literature [2] are 81.60 %, 62.80 % and 58.00 %, so this cluster validity classification method based on geometry probability is superior to the methods in literature [2].

Table 1. Classification results of random 250 pixels using method of literature.

Class	Resident Area	Shadow	Turbid Water	Dense Vegetation	Sparse Vegetation	Clear Water	Dry Land
Resident Area	**40**	0	7	0	6	0	3
Shadow	0	**1**	0	1	0	0	0
Turbid Water	0	0	**31**	0	2	5	0
Dense Vegetation	0	1	0	**37**	3	0	0
Sparse Vegetation	7	0	2	4	**53**	0	2
Clear Water	0	0	1	0	0	**27**	0
Dry Land	2	0	0	0	0	0	**14**
Total	49	2	41	42	64	32	19

Table 2. Classification results of random 250 pixels using supervised classification.

Class	Resident Area	Shadow	Turbid Water	Dense Vegetation	Sparse Vegetation	Clear Water	Dry Land
Resident Area	**40**	0	15	3	11	3	4
Shadow	0	**1**	0	1	0	0	0
Turbid Water	0	0	**14**	0	0	1	0
Dense Vegetation	0	1	0	**19**	1	0	0
Sparse Vegetation	8	0	2	18	**52**	0	3
Clear Water	0	0	10	0	0	**29**	0
Dry Land	1	0	0	0	0	0	**12**
Total	49	2	41	41	64	33	19

Table 3. Classification results of random 250 pixels using unsupervised classification.

Class	Resident Area	Shadow	Turbid Water	Dense Vegetation	Sparse Vegetation	Clear Water	Dry Land
Resident Area	43	0	6	0	12	0	14
Shadow	0	1	0	0	0	8	0
Turbid Water	0	0	16	0	1	1	0
Dense Vegetation	0	1	0	28	6	0	0
Sparse Vegetation	6	0	2	13	43	0	1
Clear Water	0	0	17	0	2	24	0
Dry Land	0	0	0	0	0	0	4
Total	49	2	41	41	64	33	19

Table 4. Classification results of random 250 pixels using methods based on geometric probability.

Class	Resident Area	Shadow	Turbid Water	Dense Vegetation	Sparse Vegetation	Clear Water	Dry Land
Resident Area	34	0	1	0	0	0	1
Shadow	0	3	0	1	0	0	0
Turbid Water	0	0	19	1	1	1	0
Dense Vegetation	0	0	0	74	7	0	1
Sparse Vegetation	4	0	0	1	41	0	4
Clear Water	0	0	3	0	0	14	0
Dry Land	10	0	0	0	1	0	28
Total	48	3	23	77	50	15	34

Table 5. Classification accuracy of random 250 pixels using method of literature.

Class	Classification Accuracy (%)	kappa	Results Assessment
Resident Area	71.43	0.6446	Overall Accuracy=Overall Kappa Statistics=0.7742
Shadow	50.00	0.4960	
Turbid Water	81.58	0.7797	
Dense Vegetation	90.24	0.8833	
Sparse Vegetation	77.94	0.7035	
Clear Water	96.43	0.9589	
Dry Land	87.50	0.8647	

Table 6. Classification accuracy of random 250 pixels using supervised classification.

Class	Classification Accuracy (%)	kappa	Results Assessment
Resident Area	52.63	0.4108	Overall Accuracy=67.20 %, Overall Kappa Statistics=0.5936
Shadow	50.00	0.4960	
Turbid Water	93.33	0.9203	
Dense Vegetation	90.48	0.8861	
Sparse Vegetation	62.65	0.4980	
Clear Water	74.36	0.7046	
Dry Land	92.31	0.9167	

Table 7. Classification accuracy of random 250 pixels using unsupervised classification.

Class	Classification Accuracy（%）	kappa	Results Assessment
Resident Area	57.33	0.4693	Overall Accuracy=64.00 % , Overall Kappa Statistics = 0.5586
Shadow	11.11	0.1039	
Turbid Water	88.89	0.8671	
Dense Vegetation	80.00	0.7608	
Sparse Vegetation	66.15	0.5451	
Clear Water	55.81	0.4909	
Dry Land	100.00	1.0000	

Table 8. Classification accuracy of random 250 pixels using methods based on geometric probability.

Class	Accuracy (%)	kappa	Results Assessment
Resident Area	94.44	0.9312	Overall Accuracy=85.20 %, Overall Kappa Statistics = 0.8145
Shadow	75.00	0.7470	
Turbid Water	86.36	0.8498	
Dense Vegetation	90.24	0.8590	
Sparse Vegetation	82.00	0.7750	
Clear Water	82.35	0.8123	
Dry Land	71.79	0.6736	

4. Conclusions

According to the basic idea and clustering steps of geometric probability-based classification method of cluster validity, we select 1498 × 1281 pixels of Xiamen Jiulong River TM images in 2002 Winter, evaluate and compare the clustering results with the supervised classification (shown in Fig. 12) and unsupervised classification (shown in Fig. 13) results by ERDAS of the same image. Experimental results show that the classification method of cluster validity based on geometric probability is superior to the literature [2], and it is also superior to the methods of supervised classification and unsupervised classification.

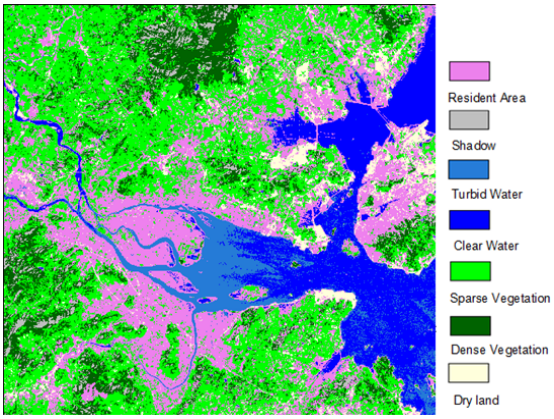


Fig. 12. Results of supervised classification.

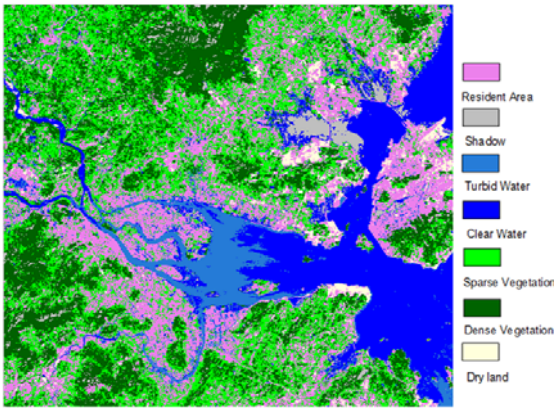


Fig. 13. Results of unsupervised classification.

Acknowledgment

This project is sponsored by the National Science Foundation of China (Grant No. 31100415) and Education Foundation of Fujian Province (Grant No. Jb12210).

References

[1]. X. W. Li, Z. Y. Mao, J. W. Li, A cluster validity function based on geometric probability, *Journal of Image and Graphics*, Vol. 13, Issue 12, 2008, pp. 2351-2356.

[2]. L. W. Huang, Z. Y. Mao, W. Z. Li, X. Q. Wang, Sheng Wu, The cluster analysis approaches based on geometric probability and its application in the classification of remotely sensed images, *Journal of Image and Graphics*, Vol. 12, Issue 4, 2007, pp. 633-610.

[3]. X. B. Gao, Fuzzy cluster analysis and its application, *Xi'an Electronic Science and Technology University Press*, 2004.

[4]. Q. He, Research of fuzzy clustering theory and applications, *Fuzzy Systems and Mathematics*, Vol. 12, Issue 2, 1998, pp. 89-94.

[5]. X. B. Gao, W. X. Xie, The research progress on the Development and application of fuzzy clustering theory, *Chinese Science Bulletin*, Vol. 44, Issue 21, 1999, pp. 2241–2251.

[6]. A. K. Jain, P. J. Flynn, Image segmentation using clustering, in: N. Ahuja, K. Bowyer, eds. Advances in image understanding: A festchrift for Azriel Rosenfeld, *IEEE Press*, Piscataway, 1996, pp. 65-83.

[7]. I. Cades, P. Smyth, H. Mannila, Probabilistic modeling of transactional data with applications to profiling, visualization and prediction, in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2001)*, San Francisco, 2001, pp. 37-46.

[8]. A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: A review, *ACM Computing Surveys*, Vol. 31. Issue 3, 1999, pp. 264-323.

[9]. R. Gelbard, O. Goldman, I. Spiegler, Investigating diversity of clustering methods: An empirical comparison, *Data & Knowledge Engineering*, Vol. 63, Issue 1, 2007, pp. 155-166.