AN ADAPTIVE CLUSTERING ALGORITHM BASED ON THE POSSIBILITY CLUSTERING AND ISODATA FOR MULTISPECTRAL IMAGE CLASSIFICATION

Wenping Liu, Chih-Cheng Hung!,[1]*, Bor-Chen Kuo# , Tommy Coleman+

Department of Digital Media, School of Information, Beijing Forestry University, Beijing, China
Department of Computer Science, North China University of Technology, Beijing, - wendyl@vip.163.com
! School of Computing and Software Engineering
Southern Polytechnic State University, Marietta, GA 30060, - chung@spsu.edu
Mathematics Education Department, National Taichung University, Taichung, Taiwan, China.
+ HSCaRS, Alabama A&M University, Normal, AL 35762

Commission VII, WG VII/4

KEY WORDS – ISODATA, Possibility Clustering, Fuzzy Clustering

ABSTRACT

For a clustering algorithm, the number of clusters is a key parameter since it is directly related to the number of homogenous regions in the given image. Although ISODATA clustering algorithm can determine the number of clusters and cluster centers dynamically, it is challenging to specify so many parameters. The possibility clustering provides the memberships that are interpreted as the degrees of possibility. So the memberships in one class are not related to the memberships in other classes. By taking the advantages of these two clustering algorithms, a new fuzzy clustering algorithm is proposed by combining the possibility clustering and ISODATA clustering algorithm. This new algorithm not only can determine the number of clusters dynamically with the degree of possibility of each date point, but also can reduce the number of input parameters of ISODATA algorithm. The splitting and merging process is evaluated by the possibility distance.

1. INTRODUCTION

Clustering is a useful approach in many different applications such as pattern recognition, pattern classification, image segmentation and data compression. It is especially popular in the unsupervised pattern recognition [1]. The objective of clustering is to classify the unlabeled data points into meaningful groups (clusters). There exist many clustering algorithms such as the K-means, fuzzy C-means and so on.

A variation of the K-means clustering algorithm, called ISODATA clustering, uses splitting and merging clusters methods to do the clustering [1]. Using this variant, the optimal partition starting from any arbitrary initial partition can be obtained [2]. However, it requires many parameters to be specified.

One of the most popular fuzzy clustering algorithms is called the fuzzy C-means (FCM) algorithm. The probability constraint is used in the FCM so that the membership of a data point over all clusters must sum to 1 [3]. Because of this constraint, the generated memberships are related to each other in the different classes. Therefore, the memberships do not correspond to the degree of a cluster belonging intuitively. The possibility clustering algorithm derived from the FCM drops this constraint [3]. The memberships with the possibility framework are interpreted typically in terms of the degree of belonging.

In order to use the advantages in the possibility clustering and traditional "hard" clustering, a new fuzzy clustering algorithm, possibility clustering combined with ISODATA, is proposed. This new clustering not only reduces the number of input parameters in ISODATA, but also makes the data membership to interpret the degree of possibility of the data point belonging to the classes.

The paper is organized as follows. Section 2 introduces ISODATA clustering algorithm. Section 3 describes the possibility clustering. The new proposed adaptive possibility clustering with ISODATA is explained in Section 4. Experimental results are presented in Section 5. The conclusion is given in Section 6.

2. THE ISODATA CLUSTERING ALGORITHM

ISODATA algorithm is an unsupervised classification [1, 5, 6]. It is similar in principle to the K-means algorithm. However, the ISODATA algorithm determines the number of clusters dynamically. To run the ISODATA algorithm, a lot of parameters such as initial cluster means, splitting parameters, lumping parameters, the minimum number of pixels in a cluster and the number of iterations must be specified. Once these parameters are defined, each sample of the feature space is grouped to the nearest cluster center. The total number of grouped samples in each cluster must meet the minimum required amount. The cluster is eliminated if the minimum amount cannot be reached. After that, compute the mean of the rest grouped samples to update each cluster center. If the splitting condition is met, split that cluster into two clusters. If

*Corresponding author.

This work was supported by the Beijing Municipal Education Commission, China (No. KM200710009005) and "863" Hi-tech research and development program of China(No. 2006AA10Z232)

the lumping condition is met, lump the two clusters. After either splitting or lumping, the new cluster centers have to be recomputed by newly grouped samples. The algorithm terminates until the maximum number of iterations is exceeded or the converged status of cluster centers occurs. Data does not need to be normally distributed. If the enough iteration is applied, this clustering algorithm is easy to find the true clusters within the data. However, more computation time is needed.

3. THE POSSIBILITY CLUSTERING ALGORITHM

Besides the ISODATA clustering algorithm, the training stage of multispectal image classification can also be achieved by other fuzzy clustering algorithms such as possibility clustering. Most fuzzy clustering algorithms are derived from Bezdek's fuzzy C-means (FCM) algorithm [4]. In the possibility clustering, the data classification result can be interpreted as a possibility classification. The membership values represent the possibility degree of the points belonging to the classes [3].

In the FCM, the probability constraint forces a data point over all classes sum to 1 [3]. This constraint is to avoid the trivial solution, which is obtained from the objective function, of all the memberships being to 0. Unlike the FCM, in the possibility clustering the membership of a data of one cluster is not relative to other clusters, but only depends on the distance of the data in respect of the cluster prototype. To change the probability constraint in the possibility clustering algorithm, an additional term is added to the FCM objective function. The new objective function makes the memberships of the representative points of clusters to be as high as possible, while the unrepresentative points of clusters to be as low as possible. This makes the membership have more typical interpretation and degree compatibility [3, 7]. So the objective function becomes

$$J(\theta,U)=\sum_{i=1}^{N}\sum_{j=1}^{C}u_{ij}^{q}d^{2}\left(x_i,\theta_j\right)+\sum_{j=1}^{C}\eta_j\sum_{i=1}^{N}(1-u_{ij})^{q} \tag{1}$$

where θ represents a list of cluster centers, U represents a list of memberships, u_{ij} is the grade of membership of the ith data point x_i in the jth cluster, $d(x_i,\theta_j)$ is the distance between the ith data point x_i and the jth cluster center θ_j, N is the total number of feature points, C is the number of classes and q is a weighting exponent called a fuzzifier. Here, η_j are suitably chosen positive constants [3]

The minimization $J(\theta,U)$ with respect to u_{ij} leads to the equation

$$u_{ij}=\frac{1}{1+\left(\dfrac{d^{2}(x_i,\theta_j)}{\eta_j}\right)^{\frac{1}{q-1}}} \tag{2}$$

where $i=1,2,...,N$, $j=1,2,...,C$. So the grade of membership value u_{ij} exclusively depends on the distance of a data point x_i to jth cluster center. The updated u_{ij} value has no relation to the other clusters [3].

In general, the value of η_j can be obtained by running the generalized fuzzy algorithmic scheme (GFAS) [3]. After the GFAS converges, η_j is given by

$$\eta_j=\frac{\sum_{i=1}^{N}u_{ij}^{q}d^{2}(x_i,\theta_j)}{\sum_{i=1}^{N}u_{ij}^{q}}. \tag{3}$$

The fuzzifier, q, determines the fuzziness of the final possibilistic C-partition and the shape of the possibility distribution [3]. The greater the q is, the fuzzier the memberships are. The value of q is usually set to 2.

Since the second term in the objective function equation (1) is independent of the cluster centers and the distances between the data points and the cluster centers, it is clear to conclude that updating each cluster center is the same as that of the FCM by using the following equation [3]

$$\theta_j=\frac{\sum_{i=1}^{N}u_{ij}^{q}x_i}{\sum_{i=1}^{N}u_{ij}^{q}}. \tag{4}$$

The possibility clustering algorithm is outlined in the following steps.
Step 1: Choose the number of clusters, initialize cluster centers $\theta_j(0)$, $j=1,2,...,C$, set count t to 0.
Step 2: Run the generalized fuzzy algorithmic scheme (GFAS) to get fixed η_j using equation (3).
Step 3: Based on $\theta_j(t)$, compute the possibility memberships $u_{ij}(t)$ using equation (2) for all the samples in different clusters. The value of q is defined as value 2.
Step 4: Based on $u_{ij}(t)$, update all cluster centers $\theta_j(t+1)$ using equation (4).
Step 5: If $\|\theta(t+1)-\theta(t)\|<\varepsilon$, where ε is a small constant, the iteration terminates. The final result of new cluster centers is accepted. Otherwise, increment count t by 1 and go to Step 3.

4. THE ADAPTIVE POSSIBILITY CLUSTERING WITH ISODATA ALGORITHM

Although ISODATA clustering algorithm can determine the number of clusters and cluster centers dynamically in the training stage, it is challenging to specify so many parameters to run the ISODATA. The possibility clustering provides the memberships that are interpreted as the degrees of possibility. So the memberships in one class are not related the memberships in other classes. That can solve the "equally likely" and "unknown" problems of data points more properly [3]. By taking the advantages of these two clustering algorithms, a new fuzzy clustering algorithm is proposed by combining the possibility clustering and ISODATA clustering algorithm. This new algorithm not only can determine the number of clusters dynamically with the degree of possibility of each date point, but also can reduce the number of input parameters of ISODATA algorithm. The splitting and merging process is evaluated by the possibility distance.

The proposed algorithm is outlined as follows.

Step 1: Randomly pick some data points as the initial cluster centers. Specify the following parameters: the maximum total number of clusters, minimum total number of clusters and number of iterations allowed.

Step 2: Based on the current cluster centers, run generalized fuzzy algorithm scheme (GFAS) to determine the η_j for each cluster S_j by using equation (3).

Step 3: Update the membership u_{ij} using equation (2).

Step 4: Update each cluster center θ_j using equation (4).

Step 5: Compute the membership u_{ij} of each data point in the new cluster centers by using equation (2).

Step 6: Based on the current cluster centers, distribute the data points using the following rule: Compare the membership u_{ij} of data point x_i in all the clusters, x_i is assigned to the cluster S_j with the largest u_{ij} value of data point x_i. That is $x_i \in S_j$ if $u_{ij} > u_{ik}$, $k = 1,2,\ldots C$ and $j \neq k$.

Step 7: (a) If this is the last iteration, go to Step 18

(b) If the current total number of clusters is less than the minimum total number of clusters, go to step 14 to split cluster S_j.

(c) If the current total number of clusters is greater than the maximum total number of clusters, go to Step 8 to merge.

(d) Otherwise, go to Step 12.

Step 8: For each data point x_i, find the cluster pair which has two largest memberships of data point x_i in two different clusters. The value of these two largest memberships must be no less than 0.6, respectively. If cluster pair can be found of data point x_i, that means x_i is in these two clusters at the same time with high possibility. If no cluster pair is found for all the data points, go to Step 2 and increase iteration number by 1.

Step 9: Count the number of data points that are in both clusters for each cluster pair.

Step 10: Repeat the cluster pair merging according to the following rule: Merge two clusters in a cluster pair that has the largest number of overlapped data points. The total number of overlapped data points must be at least 20 percent of the cluster that with more data points. Then update the cluster centers as described below. Suppose these two clusters in a cluster pair have cluster center θ_k and θ_l respectively. If neither θ_k nor θ_l has been used for merging in this iteration, merge these two clusters and get the new cluster center using the following relation

$$\theta_{new} = \frac{1}{N_k + N_l}(N_k\theta_k + N_l\theta_l)$$

(5)

where N_k, N_l is the number of data points in cluster S_k, S_l respectively. Delete θ_k and θ_l, and reduce the N_c by 1 where is N_c current number of clusters.

Step 11: If no more cluster pairs need to be merged, go to Step 2 and increase iteration by 1.

Step 12: Compute the average membership $\bar{u}_j$ of data points X in each cluster domain S_j using the relation

$$\bar{u}_j = \frac{1}{N_j}\sum_{X\in S_j} u_{ij} \quad , \; j = 1,2,\ldots C \qquad \ldots$$

(6)

Step 13: If the average membership $\bar{u}_j < 0.5$, $j = 1,2,\ldots C$, go to step 14 to do splitting of corresponding cluster S_j. Otherwise go to step 2 and increase iteration by 1.

Step 14: Calculate standard deviation of the cluster S_j along the kth dimension (kth bands), using the relation:

$$\sigma_{jk} = \sqrt{\frac{1}{N_j}\sum_{X\in S_j}(x_{ik} - \theta_{jk})^2} \quad , \qquad k = 1,2,\ldots n$$

(7)

where n is the sample dimensionality (number of bands), x_{ik} is the kth component of ith data point in cluster S_j (that means the ith data point of the kth band in cluster S_j), θ_{jk} the kth component of cluster center θ_j (that means the cluster θ_j in the kth band), N_j is the number of data points in cluster S_j.

Step 15: Find the maximum standard deviation $\sigma_{j\max}$ among all the components of cluster S_j.

Step 16: Split cluster S_j into two new cluster centers θ_j^+ and θ_j^-, delete cluster center θ_j. Cluster center θ_j^+ is formed by adding a given value γ_j to the component of cluster center θ_j which corresponds to the component with the maximum standard deviation among. Cluster center θ_j^- is formed by subtracting γ_j from the same component of cluster center θ_j. The value of γ_j can be specified by using relation $\gamma_j = k\sigma_{j\max}$, where $0 < k \le 1$. Here, we choose $k = 0.5$. Then delete cluster center θ_j, increase N_c by 1.

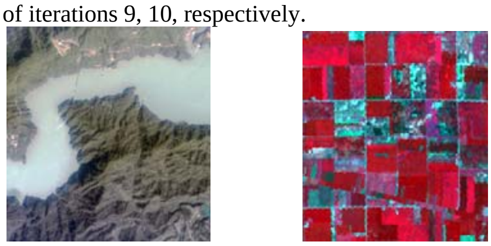
Step 17: Go to step 2 when no more splitting is needed and increase iteration by 1.

Step 18: The algorithm is terminated.

5. EXPERIMENTS

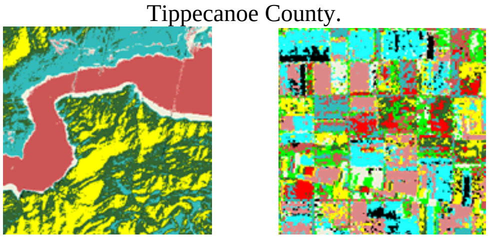
Experiments were performed by using some satellite images to demonstrate the effectiveness of the ISODATA algorithm, the possibility clustering and the new proposed clustering algorithm. The commercial software is also used to generate classification result with ISODATA clustering method.

The original satellite images, mountain/river/village and a small portion of Thematic Map of Tippecanoe County, are shown in Figure 1. Classified results with the ISODATA algorithm, the possibility algorithm and the new proposed clustering algorithm are shown in Figure 2 (a)(b), (c)(d) and (e)(f), respectively. (a) & (b) classified results with number of initial cluster 3, 6; maximum number of clusters 10, 12; minimum pixels in cluster 10, 10; maximum clusters for lumping 6, 6; maximum iterations 15, 20; respectively. (All classified results with Euclidean Distance for lumping; lumping distance zero; sampling percentage 1.00% (2621 pixels); splitting factor 12; splitting fraction 0.50, seed method using random placement, seed size 5; interrupt interval 2). (c) & (d) the desired number of clusters 4, 7, respectively. (e) & (f) minimum number of clusters 3, 3 maximum number of clusters 6, 11, maximum number of iterations 9, 10, respectively.

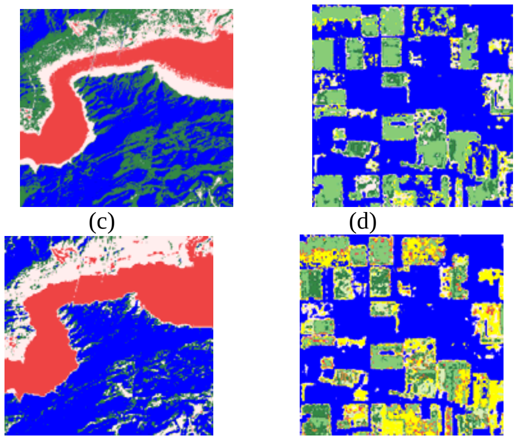


(a) (b)

Figure 1: Original satellite images. (a) mountain/river/village satellite image, and (b) a small portion of Thematic Map of Tippecanoe County.



(a) (b)



(c) (d)

(e) (f)

Figure 2: Classification results of satellite images with ISODATA clustering algorithm (a)(b), possibility clustering algorithm (c)(d) and the proposed clustering algorithm(e)(f), respectively.

6. CONCLUSIONS

The objective function in the possibility clustering affects the degree of membership for each data point. So the memberships of feature vectors do not necessarily depend on each other in the different clusters. ISODATA clustering uses splitting and merging to obtain the optimal partition. The new proposed adaptive possibility clustering approach not only uses the typical interpretation of membership value with degree of possibility, but also uses the possibility membership as the basis to do splitting and merging. This new approach gives us a new version of fuzzy clustering which can be used for unsupervised training and other applications.

REFERENCES

[1] J. T. Tou and R. C. Gonzalez, "Pattern Recognition Principles," Addison-Wesley Publishing Co., Reading, Massachusetts, 1974.

[2] A.K. Jain and M.N. Murty and P.J. Flynn, "Data Clustering: A review," ACM Computing Surveys, vol. 31, No. 3, September 1999.

[3] R. Krishnapuram and J. M. Keller, "A Possibilistic Approach to Clustering," IEEE Transaction on Fuzzy Systems, vol. 1, No. 2, May 1993.

[4] A. Kandel, "Fuzzy Techniques in Pattern Recognition," John Wiley & Sons, Inc, 1982.

[5] E. Piazza, "Comparison of different classification algorithms of NOAAAVHRR images," Proc. SPIE, July 2000.

[6] A. Ryberg and G. Borgefors, "Integrated Method for Boundary Delineation of Agricultural Fields in Multispectral Satellite Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 39, No. 11, November 2001.

[7] L.A. Zadeh, "Fuzzy Sets as a Basis for a Theory of Possibility," Fuzzy Sets and Systems, vol. 1, pp. 3-28, 1978.