JOURNAL OF PURE AND APPLIED MICROBIOLOGY, April 2013. Vol. 7(Spl. Edn.), p. 417-423

Direct Segmentation Algorithm Research for 3D Medical Data Field

Li Xinwu

Electronic Business Department, Jiangxi University of Finance and Economics, Jiangxi, China.

(Received: 03 March 2013; accepted: 14 April 2013)

Direct 3D volume segmentation is one of the difficult and hot research fields in 3D medical data field processing. Using K-means clustering techniques, a new clustering segmentation algorithm is presented. Firstly, according to the physical means of the medical data, the data field is preprocessed to speed up succeed processing; Secondly, based on analyzing the limitation of the original K-means algorithm, the paper improves the principle of the K-means, the selection of initial cluster centers and algorithm flow of K-means cluster algorithm to improve efficiency and stability of original K_means algorithm; Thirdly, based on physical characteristics of medical 3D volume segmentation, a new pixel processing method and operational principle are designed in the improved K-means segmentation algorithm to improve segmentation accuracy and speed; Finally, the experimental results show that the algorithm has high segmentation accuracy and can improve process stability and segmentation speed greatly when used to segment 3D medical data field directly.

Key words: 3D medical data field processing, Direct Volume segmentation, K-means clustering, initial cluster centers.

The segmentation of 3D medical data field has always been an extremely challenging subject due to imaging principle, fuzzy tissue and other factors. In the past more than 20 years, people had addressed a large number of segmentation algorithms. However, the complexity of human body structure, irregularity of tissue organs as well as difference among different individuals makes the segmentation of medical data field no common theory so far. Although the segmentation of 3D medical data field is very difficult, it is one of the key technologies for data field processing and system analysis and understanding, and an extremely important step of data field visualization. Only can accurate segmentation of the data field obtain reasonable models for subsequent renderings. It can be said that the realization of 3D visualization of the medical data field is to carry on correct and reasonable segmentation(Cheng *et al*., 2007) of the image data at first.(Haralick *et al*., 2010; Bradley *et al*., 2011)

Haralick [1] and Shapira regard image segmentation as a clustering process. The clustering means mathematically that a large number of d -dimensional data samples (n units) are clustered into k classes ($k << n$) so as to maximize the similarity of the samples in same class and minimize the similarity of the samples in different classes. The clustering process is to make containing several attributes by the clustering continuous classification of the data objects algorithm automatically, and the data is cut into several classes by the identification of the data characteristics. Therefore, the algorithms can be explored by the clustering rules and the clustering

* To whom all correspondence should be addressed.
Tel.: 0086-18970869647;
E-mail:liyue7511@163.com

basis of various targets found, and then based on which, the image is identified and segmente((Tang *et al*., 2004; Jioa *et al*., 2007).

K-means algorithm is a basic division method in the clustering methods and has better scalability, so it is widely used. In addition, while this algorithm that takes error square and criterion function as the clustering criterion function involves the clustering result into local solution easily and make it dependent on the initial value, a large amount of 3D medical image data causes bad algorithm timeliness. Moreover When K-means algorithm is used to cluster data, the stability of the clustering results is still not good enough, sometimes, the clustering effect is very good (when the data distribution is convex-shaped or spherical), while sometimes, the clustering results have obvious deviation and errors, which lies in the data analysis. It is unavoidable for the clustered data to have isolated points, referring to the situation that a few data deviate from the high-dense data intensive zone. The clustering mean point (geometrical central point of all data in the category) is used as a new clustering seed for the K-means clustering calculation to carry out the next turn of clustering calculation, while under such a situation, the new clustering seed might deviate from the true data intensive zone and further cause the deviation of the clustering results. Therefore, it is found that using K-means algorithm to process the data of isolated points has a great limitation(Tang *et al*., 2004; Jioa *et al*., 2007; Yi , 2010; Zhang *et al*., 2009; Dai *et al*., 2012).

In order to overcome these K-means algorithm shortcomings, a newly improved K-means algorithm of medical image volume segmentation is provided on the basis of improving the principle of the algorithm, selection of initial cluster centers and algorithm flow in the paper(Mohamed *et al*., 2012).

Data field pretrement

Each voxel's gray value (or color value) is given according to the people's habits or the users' requirements, and not owned by substances. Therefore, the difference of adjacent data has a certain meaning while the absolute value of each data is of no importance in the data field. This algorithm suggests that the function values of original 3D data field is within the range of 0-255 in the normalization integration, and processed data replaces original data to give the gray level value so as to provide a gray level field for the feature extraction, decrease the post-treatment memory demand and improve the post-treatment speed.

Although it is not simple for such problems as the time consumption that the data field turns 16-bit and 12-bit gray level images into 8-bit under the premise of keeping up the key information of the images in the process of the normalization pretreatment, the process is over in data format conversion of the pretreatment and any data field will be treated only once so as not to affect the efficiency of the whole algorithm.

K-Means algorithm thinking

Steps for K-means clustering algorithm are (Mohamed *et al*., 2012; Yang, 2012) (see Figure 1).

(1) Select n objects as the initial cluster seeds on principle;
(2) Repeat (3) and (4) until no change in each cluster;
(3) Reassign each object to the most similar cluster in terms of the value of the cluster seeds;
(4) Update the cluster seeds, i.e., recompute the mean value of the object in each cluster, and take the mean value points of the objects as new cluster seeds.

Improvement of k-means algorithm

Limitation of initial k-means algorithm

When K-means algorithm is used to cluster data, the stability of the clustering results is still not good enough, sometimes, the clustering effect is very good (when the data distribution is convex-shaped or spherical), while sometimes, the clustering results have obvious deviation and errors, which lies in the data analysis. It is unavoidable for the clustered data to have isolated points, referring to the situation that a few data deviate from the high-dense data intensive zone. The clustering mean point (geometrical central point of all data in the category) is used as a new clustering seed for the K-means clustering calculation to carry out the next turn of clustering calculation, while under such a situation, the new clustering seed might deviate from the true data intensive zone and further cause the deviation of the clustering results. Therefore, it is found that using K-means algorithm to process the data of

isolated points has a great limitation(Jioa *et al*., 2007; Zhang *et al*., 2009; Manjunath *et al*., 2011)

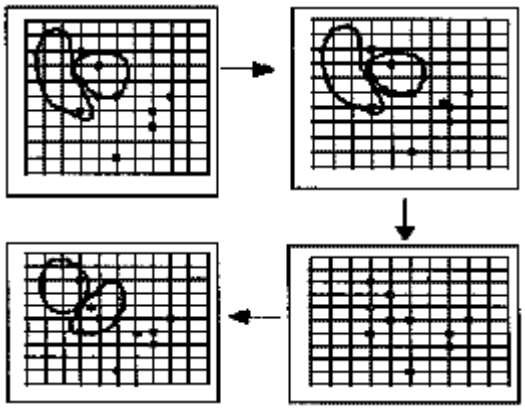

Fig. 1. K-means algorithm procedures

Improving k-means principle

When calculating the K turn of clustering seeds with the improved algorithm, those data in the cluster having a great similarity to the K-1 category seeds should be adopted to calculate their mean points (geometrical center) as the clustering seed of the K turn and the specific calculation method is below:

(1) For the cluster $c_{i(k-1)}$ obtained through the K-1 turn of clustering, the minimum similarity $sim\_\min_{i(k-1)}$ of the data in the cluster to the clustering seed $s_{i(k-1)}$ of the cluster is calculated;

(2) The data in the cluster $c_{i(k-1)}$ is calculated that has a similarity of more than $1-\beta*(1-sim\_\min_{i(k-1)})$ to the clustering seed $s_{i(k-1)}$ (among, β is a constant between 0-1), and the data set is recorded as $cn_{i(k-1)}$;

(3) The mean points of the data in $cn_{i(k-1)}$ are calculated as the clustering seed of the K turn.

a. Cluster i of the K-1 turn and its seeds

b. Cluster i of the K turn and the new seeds (initial algorithm)

c. Cluster i of the K turn and the new seeds (improved algorithm)

As seen in Figure. 2, the new clustering seeds are obviously moving toward the data intensive zone. The improved algorithm could achieve a good clustering effect on the text sets containing isolated points. For the processing of big text sets, this improved algorithm, as same as k-means algorithm, is relatively flexible and high-effective. Its time complexity is $O(nkt)$, of which, n is the number of all objects, k is the number of the clusters, while t is the iteration number of the algorithm, and generally, $k\langle\langle n$ and $t\langle\langle n$.

Improving selection of initial cluster center

Euclidean Distance has a very intuitive significance for clustering, thus, it is used in the paper to express the distance between the sample points, and the distance between Sample $X=(x_1,x_2,...,x_n)$ and $Y=(y_1,y_2,...,y_n)$ is calculated according to Formula (1).

$$d(X,Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+...+(x_n-y_n)^2} \quad ...(1)$$

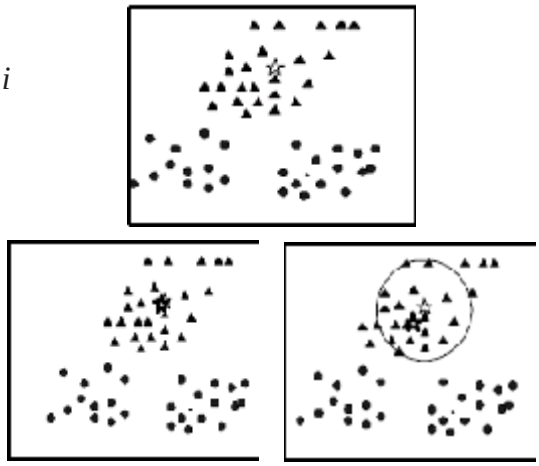

Fig. 2. The Comparison principle of K-means algorithm and the Improved Algorithm

Indication of the symbols in figure 2.

Δ means Data point in the cluster

´ means seed of the cluster i in the K-1 turn

´ means new seed the cluster i in the K turn

´ means other data points

´ % means the points within its range are used to calculate the new seeds.

The distance between a sample point and

a sample set is defined as the nearest distance between the sample point and all sample points in the sample set. The distance of a sample point and a sample set V is defied as Formula (2).

$$d(X,V) = \min(d(X,Y), Y \in V) \quad ...(2)$$

Proposed that the sample set U has n samples, clustered as category k, and the initial value of m is 1, and the improved algorithm is described as below.

(1) The distance $d(X,Y)$ between any two samples is calculated, the nearest two points in the set U are found and compose the set $Am(1 \le m \le k)$, and then these two points are deleted from the set U.

(2) The point is found that is nearest to the set Am, added into the set Am and deleted from the set U.

(3) Step 2 is repeated until the number of the sample points in the set is not less than $\alpha_{n/k}(0\langle\alpha \le 1)$, and the value of α varies from the experimental data. If the value of α is too small, it might have several initial cluster center points obtained in the same region, while if too big, it might have the initial cluster centers deviating from the intensive zone, thus, based on the experimental situation, it is suitable to value it as 0.75.

(4) If $m\langle k$, $m = m + 1$, two points nearest to the set U are found to form a new set $Am(1 \le m \le k)$ and are deleted from the set U, then the Step 2 is repeated.

(5) The sample points in k sets formed eventually are calculated for their means, so as to form k initial cluster centers.

(6) Based on k initial cluster centers, K-means clustering algorithm is used to form the final clustering.

Improving the algorithm flow

The general K-means algorithm is a gradient ascent iteration algorithm, each time of iteration could cause the corresponding increase of the target function values, and the iteration might be ended in the limited steps. However, such an algorithm also has some disadvantages, for example, the algorithm is easily trapped in the local maximum solution and such a solution depends on the selection of initial partition. Therefore, the means algorithm is used as local searching process to be inlaid in the local search structure of the iteration in order to obtain better text clustering results through the relationship between balancing the reinforcement of the local search and extending the searching range.

In the image clustering problems, D neighborhood of a partition refers to the partition obtained through randomly selecting D different texts in a certain partition and redistributing them into other clusters. In other words, the neighborhood of the current partition means the partition obtained through randomly selecting one text and redistributing it into other cluster. The calculation flow of the K-means-based iteration clustering algorithm of local searching texts is X displayed in the following.

Input: the number k of the results' clusters, containing the data set of N texts.

Output: k clusters, ensuring that the texts in all clusters are similar or correlated.

Step 1 Randomly select an initial partition $P_i = \{C_1, C_2, \ldots C_k\}$ and calculate the corresponding concept vector $c(C_i), i = 1,2,\ldots k$, then initialize the current maximum target function value $fopti$ and determine the ending conditions of the algorithm, the parameter value $\varepsilon(\varepsilon \succ 0)$ receiving the conditions and the maximum iterating times n that the target function value is not improved any more.

Step 2 Repeat.

Step 3 Perform the local search on P_k with the means text clustering algorithm to obtain a local maximum target function value $fopt$ and its corresponding partition P_k^*.

Step 4 If $fopt \succ fopti$, the current best partition

is $P_k^{'} = P_k^{*}$, $fopti = fopt$, the current partition is not improved any more, and the iteration times $t := 0$.

Step 5 Repeat

Step 6 Randomly generate a text $x_i (i = 1,2,...N)$ and repeat the following processes:

(1) If x_i is beyond the tabu list, it will be redistributed into other cluster to calculate the increase Δf of the target function value and the times of iteration without improvement is $t : t = t + 1$; while if x_i is in the tabu list, Step 6 will be repeated.

(2) If $\Delta f \succ \varepsilon$, P_k is the partition of redistribution, the target function value is $f_{opt} = f_{opt} + \Delta f$, is added into the tabu list, and the tabu length of other tabu objects is deducted 1.

(3) If $fopt \succ fopti$, $fopti = fopt$, $P_k^{'} = P_k$, and the times of iteration is $t := 0$.

(4) If x_i is tested throughout all clusters and the times of iteration without improvement is $t \prec n$, Step 6 will be repeated.

Step 7 "Until $t = n$" means there is no improved partition generated in the successive n times of iteration.

Step 8 Randomly select several texts from P_k and redistribute them into other clusters to obtain the new partition P_k.

Step 9 Until the ending conditions are met.

Pixel processing and operational princial of the improved algorithm

In order to quicken the efficiency and the ability of the algorithm to process the large-scale data field, the pixel operation and processing have to comply with the several following principles:

(1) Pre-segment the data field in the phase of the data field segmentation pretreatment, i.e. use the methods of manual interaction and model guidance.

(2) According to prior knowledge of the structure shape and the position that medical data field dissected the tissue, give the interactive definition to several seed points and take these seed points as the initial samples.

(3) According to the probability distribution of an established characteristic, directly classify the pixel points that the selected obvious characteristic belonged to a seed point, namely the points that have the obvious characteristics and definitely belong to a class will be marked as a class directly, and not be calculated.

(4) Calculate the points that have no obvious characteristics and are classified strictly through mathematical algorithm, namely the points that are possible to belong to different classes only for the edge region or the edge transitional region carry on the algorithm operation and segmentation.

(5) As for these points to be calculated, use the dot interlaced sampling to carry on the sampling calculation in the space. Namely make the sampling to calculate whether a point belongs to A class from the surrounded seed point A. When the calculation of the n point finishes, next point to be calculated will be selected as n+2 but not n+1 according to the space order if the n point belongs to A class; if the calculation result is that the n+2 point belongs to A class, the n+1 point will be fallen into A class directly; if the calculation result is that the n+2 point does not belong to A class, the attribute that the n+1 point directs towards A class will be calculated repeatedly.

x_i

Such this reduces the blindness of the defined initial sample points greatly so as to enhance the accuracy of the segmentation, and also reduces the data quantity calculated by the algorithm greatly to enhance the algorithm efficiency.

Experimental verification

This improved algorithm of the paper is realized on PC with 3D medical MRI data field. PC configurations are P4 1.8G CPU and 512M memory. In this paper, the algorithm is simulated in the utilization of the data sets with different sizes and various distributions. Due to the limited space, the following only introduces the segmentation results of actual data fields measurement for two groups

of complex medical organization MRI data field. Both of the date field size are $128\times128\times128$. The experiment results show that the segmentation effect comparison between general K-means algorithm[5] and the improved K-means algorithm presented by the paper is shown in Figure. 3 and Figure 4. Figure 3 (a) and Figure 4(a) are original image, Figure 3 (b) and Figure 4(b) are the image clustered and segmented by general K-means algorithm which lacks many details obviously, Figure 3 (c) and Figure 4(c) are the image segmented by the algorithm presented in this paper which has high segmentation precision and good visual effects; (2) The improved algorithm of the paper can improve calculation efficiency significantly and segmentation accuracy also , the experimental result of two groups of medical data can be seen in table 1.

Fig. 3. Original image and segmentation effect image with different algorithms (Data A)

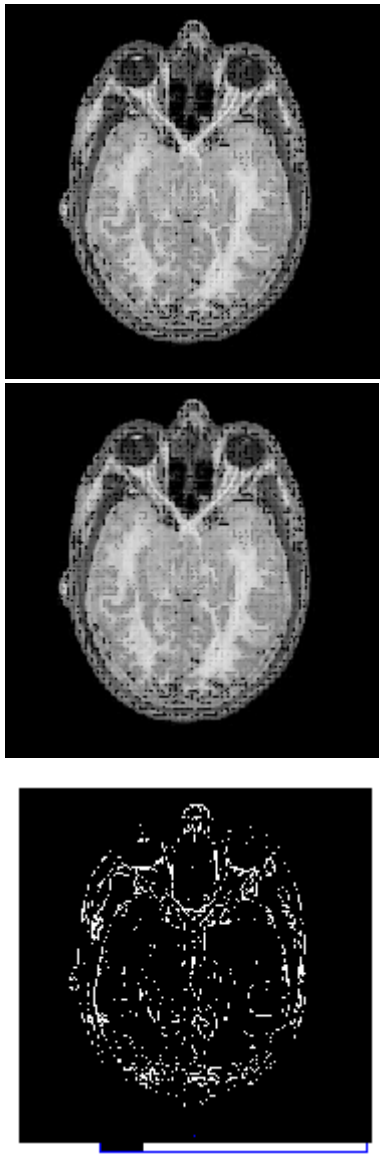

Fig. 4. Original image and segmentation effect image with different algorithms (Data B)

Table 1. Segmentation efficiency and accuracy comparison table of different algorithms

Algorithm	Improved algorithm	General K-means algorithm
Time consumption (s) of data A	18	65
Segmentation Accuracy of data A	98.9%	74.1%
Time consumption (s) of data B	20	73
Segmentation Accuracy of data B	98.4%	71.7%

In figure 3, (a) means original image of data A, (b) means Segmentation result of general K-means algorithm, (c) means segmentation result of this paper's algorithm.

In figure 4,(a) means original image of data B, (b) means Segmentation result of general K-means algorithm, (c) means segmentation result of this paper's algorith

CONCLUSION

General K-means clustering algorithm is vulnerable to the local solution in practical application, so this paper fully utilizes prior knowledge of the segmentation object to perform several pretreatments in the course of detailed computation by the thinking of seeking the optimal initial value in several samplings and one clustering as well as the improved K-means clustering algorithm. As a result, large reduction of the processing units and great improvement of the algorithm anti-interference make the algorithm improve not merely convergence speed but also segmentation accuracy. Besides, the practical application of K-means clustering segmentation algorithm is improved in 3D medical data field segmentation greatly.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China under the grant No.60963012.

REFERENCES

1. Bradley P. S., Managasa L., K-plane Clustering. *J. Opti*. 2011; **33**(1): 23-32.
2. Cheng Y, and Huang F., A Coupled Minimization Problem for Medical Image Segmentation with Priors. *Int. J. of Comp. Vis*. 2007; **38**(3):259-272.
3. Dai M, Mu J., A Fast Approach to K-means Clustering for Time Series Based on Symbolic Representation. *Int. J. Adv. Comp. Tech.* 2012; **4**(5): 233-239.
4. Haralick R. M., Shapiro L. G., Survey:Image Segmentation Techniques, *J. Com Grap.,* 2010; **64**(7): 100-132.
5. Jioa C. L., Gao .,Shi Y., Image Clustering and Segmentation Based on Improved Neural Network, *Comp. Eng. & App*., 2007; **43**(20):93-960
6. Manjunath S. S., Lalitha R., Refinement of K-means Clustering for Segmentation of Microarray Images. *J. Conv. Info. Tech*. 2011; **6**(9): 403- 411.
7. Mohamed B. Abubaker, M., K_means-Based Convex Hull Triangulation Clustering *Algorithm. Res. Notes* 2012; **9**(11):19-29.
8. Tang Y., Rong Q.S., An Implementation of Clustering Algorithm Based on K-means, *J. Hubei Univ.* 2004; **22** (1):69-71.
9. Yang X. X., Expansion Research on K-means Clustering Algorithm Based on Rough Set. *Int. J. Adv. Comp. Tech*. 2012; **4**(10): 221-227.
10. Yi S., Global Optimization for K-means. *IEEE Comp*. 2010; **19**(7): pp. 45-54.
11. Zhang Y. F., Mao J. L., An improved K-means *Algorithm, Comp.Appl*. 2009; **23**(1):31-33.