An Efficient Clustering Technique for Reassignment of Isolated Regions

S. Kami Makki†, David A. Heitbrink†, Xiaohua Jia††

† Department of Electrical Engineering & Computer Science at the University of Toledo, Toledo, Ohio, U.S.A.,
† Department of Electrical Engineering & Computer Science at the University of Toledo, Toledo, Ohio, U.S.A.,
†† Computer Science Department of the City University of Hong Kong, Kowloon, Hong Kong,

Summary
Fuzzy C-Means (FCM) clustering is a popular technique used in image segmentation and pattern recognition. However one of the main problems with FCM clustering is the lack of spatial context. That is FCM often fails with irregularly shaped clusters. This can lead to the creation of isolated regions; isolated regions are those regions that are not connected with the main body of the clusters. We propose a post-processing technique whereby these misclassified regions are identified and reassigned to their proper clusters.

Keywords:
Classification, fuzzy cluster, video game, image processing, spatial access, Flooding Isolated Region Reassignment

1. INTRODUCTION

Clustering is a process of partitioning a set of data into a set of meaningful groups which consist of data items that have similar characteristics. Therefore a cluster is a collection of data items that are similar to one another and hence can be treated collectively as a group.

Clustering techniques have been applied to a wide variety of research problems in many areas such as statistical, machine learning, data mining, image processing, information filtering and retrieval [1][5][8][11][11] and other areas such as biology, psychology, archaeology, and geography [6][7].

Due to this diverse usage, clustering techniques have different measures for classifying or grouping a set of data items. One classification method classifies the clustering method into hierarchical and partitioning clustering techniques and another classifies the clustering method into exclusive and non-exclusive clustering techniques. The non-exclusive clustering also will be further divided into overlapping, hierarchical, and probabilistic clustering techniques. In this classification method, the exclusive technique partitions a data set into a number of exclusive clusters in which each data item belongs to exactly one cluster, while the non-exclusive technique assigns a data item to several clusters. A number of different algorithms also have been developed for each of the above techniques, such as K-means, Fuzzy C-means (FCM), Hierarchical, Mixture of Gussians, Probabistic, Density-based, Grid-based algorithms. The K-means algorithm is a simple clustering algorithm which classifies a data set into k exclusive clusters. While the Fuzzy C-means algorithm partitions a data set into a number of overlapping clusters using membership function. However the Fuzzy C-means algorithm can frequently produce isolated regions. These regions do not have a connection with the main body of the clusters that they are belong to and they are connected to other clusters. We propose a solution in which these regions are identified and reassigned to their proper clusters.

In the next section we provide an introduction to fuzzy clustering algorithms and show the shortcomings of these algorithms. In section 3 we explain our solution for identifying the isolated regions, and demonstrate how the shortcomings of Fuzzy C-means have been eliminated and finally, the conclusions and a direction for future research are presented in Section 4.

2. FUZZY CLUSTERING ALGORITHMS

The Fuzzy C-Means algorithm is one of the first among fuzzy clustering techniques. It has been popular in the image segmentation field for some time. The FCM algorithm was first proposed by J. Dunn [2] in 1973. In 1981, Bezdek proposed several important enhancements to this technique [2]. The FCM technique is a non-exclusive clustering algorithm which partitions a data set in such a way that it allows one data item to belong to two or more clusters. The data items are bound to a cluster by means of a membership function. The FCM technique consists of the following steps.

1. The first step is to choose an arbitrary number of clusters, and then arbitrarily assign to each cluster a centroid V_i, where $1 \le i < k$.
2. The second step is to calculate the membership degree μ_c by using Equation 1 for each point x which exists in the set X of each cluster v_i in set V. The term m in Equation 1 is an arbitrary number greater then 1.

$$\mu_{C_i}(x) = \frac{1}{\sum_{j=1}^{k}\left(\frac{\|x - v_i\|^2}{x - v_j}\right)^{\frac{1}{m-1}}} \qquad (1)$$

3. The third step is to update the centroid for each cluster, using Equation 2.

$$v_i = \frac{\sum_{x \in X} (\mu_{C_i}(x))^m \times x}{\sum_{x \in X} (\mu_{C_i}(x))^m} \qquad (2)$$

4. The fourth step is to test for completion. The completion is when, the difference between the old centroid and the calculated centroid in step three falls below some threshold, if not repeat step two through step four until completion.

In comparison with many other clustering techniques, FCM is very straightforward to implement and is simple to understand. However there are two major problems of using FCM. The first is, FCM has no spatial context; it lacks awareness of natural gaps that exist in data sets. The second drawback to FCM is its sensitivity to noise. Smoothing filters have long been applied in preprocessing steps to reduce the noise in datasets, which causes a loss of valuable information.

Therefore the post processing techniques have been applied to adjust the membership degrees. One such a technique is vector probability diffusion [12]. It uses analysis of vectors to determine out of place vectors (vectors that are highly dissimilar to neighbor vectors) and to correct them. Others have used rule-based systems as post processing technique. One technique proposed by Pham [9] called Adaptive Fuzzy C-Means (AFCM), uses a technique to penalize data points with unusually high or low membership when compared to points surrounding the data point.

Recently Fuzzy Map Clustering (FMC) a new technique for fuzzy clustering was introduced to handle the irregular shaped continuous clusters by Fu and Medico [4]. (The term irregularly shaped is usually used to describe clusters that are long thin with lots of curves).This technique relies on the producing of a series of micro-clusters and merging the clusters until only a given number of clusters remains.

The FMC process is straightforward. The first step is to find a set of local maxima based on density. The density is determined by the number of neighbors that are within a given radius. Local maxima are points that have a density greater than its n-nearest neighbors. These points are then set as the cluster supports, which are used as prototypes for the clustering. At the same time, outliers are identified as all points that have very low densities. The next step in the process is to assign membership degrees to all data points in the set. Once the membership degrees for each point have been calculated, the next step in the process is to merge the clusters until only a desired number of clusters remain. A drawback with the FMC is, it tends to produce clusters with unbalanced sizes. That is, FMC tends to produce one large segment and a number of smaller segments.

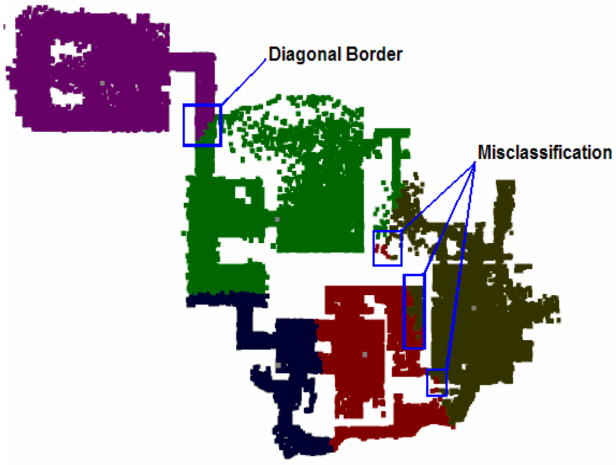


Fig. 1 – Segmented space

However the major problem of using fuzzy clustering algorithms in a number of applications such as video games is the obstruction of isolated points and segments in respect to some clusters (regions) which identified in Figure 1 as "Misclassification" points or segments and the proper handling of irregularly shaped clusters [4]. Since the isolated regions reduce the efficiency required for many applications. For example in a video game application, if a small corner of a room is in a different server's region of control, then this will increase the number of handovers and multi-server window queries. As ideally, in a proposed segmentation system, continuous regions would be considered optimal as they reduce the number of handovers that would have to be preformed. Also the implementation

Manuscript received October 16, 2006.
Manuscript revised October 25, 2006.

of the Adaptive FCM algorithm confirms that this algorithm is not well suited to correct the isolated points and segments which are caused more by the irregular size than by noise.

Given the inability of FCM and many other post-processing techniques for obstruction of isolated points/segments and handling of the irregularly shaped continuous segments we propose a new technique based on "flooding." This technique is based on an iterative technique and relies on attempting to connect members of segments, and removing data points from segments that are not connectable to the centroid of a segment.

3. The Flooding Isolated Region Reassignment

The flooding isolated region reassignment (FIRR) technique was proposed to overcome the shortcomings of the FCM process for the segmentation of a virtual world space as shown in Figure 2, for playing games in video games such as Quake II (as it is a popular game).

One of the major problems with using the FCM algorithm was its production of isolated points in respect to the cluster as detected in Figure 1. Ideally, in a proposed segmentation system, continuous regions would be considered optimal as they reduce the number of handovers that have to be made. Therefore if a small corner of a room were in a different server's region of control then this would increase the number of handovers and multi-server window queries.

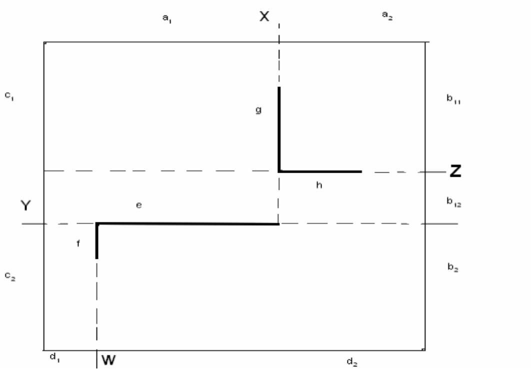


Fig. 2 – Segmented world space

One solution to this problem is to "flood" the segment to find isolated points, and reassign the points to another region. The flooding technique is based on an iterative technique of trying to connect all points that belong to a particular cluster. In this case, a point is said to belong to a cluster if its membership degree is highest to that particular cluster.

The flooding technique starts by selecting all the points within distance R of the center of gravity of the cluster as the initial starting point. All the selected points are assigned an iteration number of one; all other points are assigned zero. The current-iteration-number is set to one. The medoid may also be selected instead of the center of gravity.

The second step is to select all the points within distance R to any point with an iteration-number equal to the current-iteration-number of one and that has an iteration-number of zero. Each selected point is then assigned an iteration-number of two.

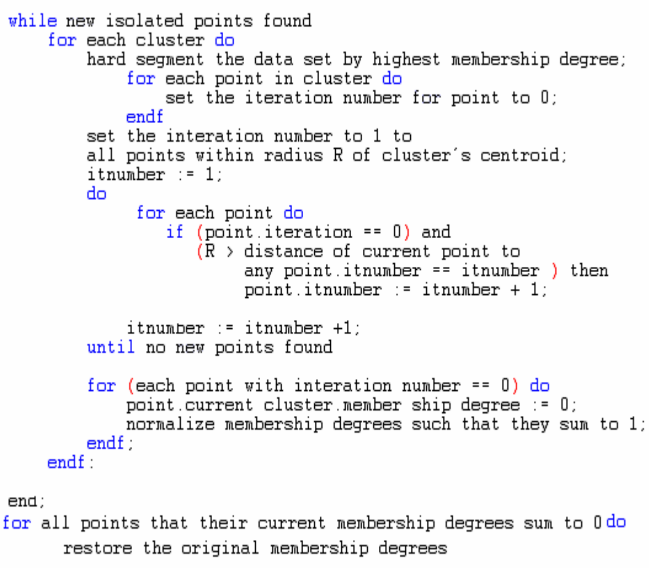
The third step is to repeat the actions of the second step and increase the current iteration-number, in other words the selected points would be assigned to three next, then to four and so on. This is done until no new points are selected.

All other points with zero as their iteration-number can be seen as isolated points. The membership degree to that cluster is then set to zero, and the rest of the membership degree's to the other clusters are normalized so that they sum to one.

At this point, the flooding technique will be repeated as the isolated points have been reassigned to new clusters. The flooding technique is repeated until no new isolated points are found. The description of the flooding algorithm is depicted in Figure 3.

The maximum number of iterations that can be performed is equal to the number of clusters that exist in the data set. This is because after each point is reassigned its membership degree is set to zero. Henceforth a point can only be reassigned once for each cluster before its membership degrees sum to zero.

Although FIRR can remove the isolated points and create a continuous segmentations to enhance the efficiency of the application. It has a number of short comings which we have addressed and the propose solutions have been provided. The first problem is the time required for computation and elimination of the isolated points. For example a worst-case scenario would require $C\sum_{i=0}^{N} i$, where C is the number of clusters and N is the number of data points, and an average case would require $I = C(\sum_{n=n-ave}^{0} (\sum_{i=n}^{n-ave} i))$, where *ave* is the average number of points connected per iteration. The above equations also do not take into account the use of spatial access or point access methods, which would reduce the number of compares to select all points within a distance of R. Therefore employing a spatial access method such as a sphere-tree could greatly increase the efficiency of the algorithm.



```
while new isolated points found
    for each cluster do
        hard segment the data set by highest membership degree;
            for each point in cluster do
                set the iteration number for point to 0;
            endf
        set the interation number to 1 to
        all points within radius R of cluster's centroid;
        itnumber := 1;
        do
             for each point do
                if (point.iteration == 0) and
                   (R > distance of current point to
                        any point.itnumber == itnumber ) then
                        point.itnumber := itnumber + 1;

            itnumber := itnumber +1;
        until no new points found

        for (each point with interation number == 0) do
            point.current cluster.member ship degree := 0;
            normalize membership degrees such that they sum to 1;
        endf;
    endf:

end;
for all points that their current membership degrees sum to 0 do
      restore the original membership degrees
```

Fig. 3 – Flooding algorithm

A second problem with FIRR is that it can produce unassigned points. There are two solutions exist for dealing with these points: they can be treated as noise and be removed from the dataset, or their original membership degrees can be restored. Should too many unassigned points be created, it may become necessary to increase the value of R (that is the distance from the center of gravity of the cluster).

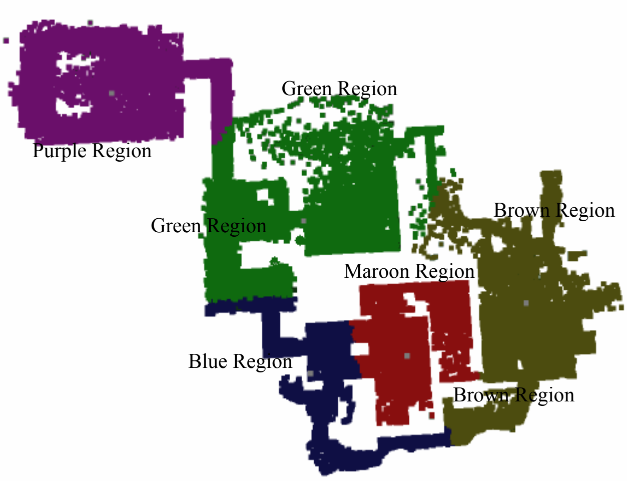


Fig. 4 - FIRR Post-Processing of Figure 1, using the restoration of isolated point's membership degrees

The next problem with the FFIR approach is the doughnut problem. Given a case where a segment took the form of a doughnut, the centroid would be at the center of the doughnut. This could cause the flooding to fail as no data points to the doughnut may be within a distance of R to the centroid. Given this case, using the medoid, or the median data point instead of the centroid, could resolve this problem. This, of course, could come up again if the medoid is an isolated point inside the middle of the doughnut. A further step could be taken to use the closest local maxima (based on density) to the medoid.

The results from the enhanced flooding isolation region reassignment (FIRR) process can be seen in Figure 4. The isolated brown segments next to the maroon region which were identified as "Misclassification" segments in Figure 1, now are connected to the rest of the brown region. That is because, as mentioned earlier these misclassified segments can cause needless handovers regardless of their size. (Note: in Figure 1 and 4, the brown color region is shown on the far right hand side, the green color region in the middle top, the blue color region at the bottom of green color and on the left hand side of the maroon region, the purple color region on the top most left hand side, and the maroon color region is shown in the middle bottom among the green, brown and blue color regions.) The EFIRR process was able to identify the isolated corners (which were incorrectly classified as part of brown region and visa versa) and reassign them to maroon region and the same for other regions as well.

Overall, the FIRR technique provides better segmentation of a virtual world for a video game application (e.g. Quake II). In this process the isolated segments are usually identified and reassigned to other regions. The cost of FIRR is similar to many other techniques that require n-nearest neighbors.

4. Conclusions & Future Research

In this paper, we have introduced the FIRR process for reclassification of misclassified regions produced by FCM due to the irregular shape of a cluster. To this end, FIRR was successful. As shown in Figures 1 and 4, the isolated brown and maroon segments were reclassified by eliminating the isolation of these segments from the main body of their respective clusters.

For the future research we intend to enhance further the FIRR process to improve its ability to handle the noise and poorly separated clusters as well. For example a metric called the strength of connection factor could be introduced into the algorithm, whereby when regions are being connected from the centriod outward, the strength of connection could be used as a threshold to determine if a connection exist between a point and the cluster. This strength of connection could be based on several metrics

Manuscript received October 16, 2006.
Manuscript revised October 25, 2006.

such as density and vector probability.

5. REFERENCES

[1] M. R. Anderberg, *Cluster Analysis for Applications*. Academic Press, Inc., New York, NY. 1973.

[2] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.

[3] J. C. Dunn, A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics 3: 32-57, 1973.

[4] Li-Min Fu, and E. Medico, "FMC, a Fuzzy Map Clustering algorithm for microarray data analysisItalian", Bio-information Technology Society 2004 Conference, March of 2004, http://bioinformatics.cribi.unipd.it/bits2004/abstracts/22.pdf

[5] A. Jain, and P. J. Flynn, *Three Dimensional Object Recognition Systems*. Elsevier Science Inc., New York, NY, Eds. 1993.

[6] A. Jain, AND R. C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ, 1988.

[7] A. Jain, M. N. Murty, and P.J. Flynn, Data clustering: a review. ACM Computing Surveys, 31, 3, 264-3231999.

[8] S. K. Makki, D. Heitbrink, X. Jia, Flooding Isolated Region Reassignment, in the Proceedings of the IEEE International Conference on Granular Computing (GrC'06), Atlanta, Georgia, 2006.

[9] Dzung L. Pham, Spatial Models for Fuzzy Clustering, Computer Vision and Image Understanding, 84(2): 285-297, 2001.

[10] E. Rasmussen, Clustering algorithms, In Information Retrieval: Data Structures and Algorithms, W. B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Inc., Upper Saddle River, NJ, 419–442, 1992.

[11] G. Salton, Developments in automatic text retrieval. Science 253, 974–980. 1991.

[12] R. Simmons, Peer-to-Peer Networking in Massively Multiplayer Online Games, Multimedia Systems Conference, 2004, http://mms.ecs.soton.ac.uk.