

# A Fuzzy Clustering Approach for Supervision of Biological Processes by Image Processing.

**J.C Atine**  
LAAS-CNRS,  
7 avenue du Col. Roche,  
31077 Toulouse.  
jcatine@laas.fr

**A. Doncescu**  
LAAS-CNRS,  
7 avenue du Col. Roche,  
31077 Toulouse.  
adoncesc@laas.fr

**J. Aguilar-Martin**  
LAAS-CNRS,  
7 avenue du Col. Roche,  
31077 Toulouse.  
aguilar@laas.fr

## Abstract

We present in this paper a new clustering method based on a fuzzy logic tree (T-LAMDA). This method has been compared with fuzzy clustering, Mean-Shift and Watershed. The application is the segmentation of color images for the diagnostic of microbial population into a bioreactor.

**Keywords :** Fuzzy logic, fuzzy decision tree, T-LAMDA, bioprocess, diagnostic.

## 1 Introduction

Before the underlying scientific principles were understood, fermentation has been known and practiced by mankind since prehistoric times. The next stage in its development was dominated by the success in the use of regulatory control mechanisms for the production of amino acids. The first breakthrough came in the late 1950s and early 1960s, when a number of Japanese researchers discovered that regulatory mutants were capable of over-producing amino acids. The higher value of final products like vitamins, baker's yeast, and antibiotics is produced by statistically non-stationary biochemical processes which need continually adaptive recipes for optimal performance. The microbiologists attempt to use computer science tools to understand and control the physiological states of these microorganisms. Different parameters directly link to microbial activity have been measured during a bio-process. We tried to identify alive cells from dead cells. For the present, some colorant is used to put in evidence the strong metabolic activity. Image processing allows to separate the colored

dead cells from not colored alive cells when a colorant is used.

A new method of color segmentation named T-LAMDA is presented in this article. Our method constructs a fuzzy decision tree using Union and Find operators. The different groups are represented by a tree labelled by the root. Each node of the fuzzy tree allows us to cluster the pixels with the same spatial and colorimetric similarity.

The fuzzy induction method we have adopted is based on a fuzzy decision tree that is an extension of binary decision trees (Breiman et al. 1984 ; Quinlan, 1986). A fuzzy decision tree is a tree with fuzzy decisions functions. A 0-1 decision tree utilize 0-1 decision functions. The decision tree Algorithms operate in the same way to construct a downward tree, from the root to the leaf, according to a general method to get arborescent system of clustering. The choice criterion in the algorithm steps is the difference between objects [1] [2]. Decision trees are simple, clear and easy to do tools. The arborescent structure is similar to a rule based "if ... then", this fact explains a decision obtained through a path. Decision tree technique is a well-known method to take classification decisions in pattern recognition. Its principal property focuses on the fact that a broad number of classes could be maintained while the time of the final decision is minimized with small local decision. The "dividing to reign" technique (divide and conquer strategy) is one of the great families of approaches to resolve problems like learning. First, it tries to identify partial problems to find a solution and then, it solves general problems combining these solutions. The decision trees learning

algorithms are built on this principle.

Fuzzy decision tree takes numeric and symbolic data into account during the tree construction and clustering process. Reasoning using numeric-symbolic values, such as size is tall or middle size, is reasoning close to human thought and the use of fuzzy set theory lead a better comprehensibility to decision tree during the process of numerical data. It has been showed that fuzzy set theory gives a better robustness when a new example is classified [3]. We choose different methods for comparison because : fuzzy c-mean allows to compare with a fuzzy clustering method ; mean shift is a method based on estimation of probability density and choosed because T-LAMDA uses the update of the membership function by using an estimator ; watershed is a very known method used in cell's segmentation.

## 2 The Learning Algorithm LAMDA

Each object is described by a set of  $n$  attributes or descriptors, and represented by vectors of  $n$  components. The set of those vectors will be called data base. Qualitative or quantitative descriptors can be considered in LAMDA method. Qualitative descriptors take their values from a non-ordered set (color, sharp). Quantitative ones take their values from a totally-ordered set, not necessarily numerical (e.g. weight, temperature, etc).

To make possible a direct confrontation between classes and objects, it is necessary that the concepts are described with the same descriptor used for observations. Given an object  $x$  and a concept  $C$ , LAMDA computes for any descriptor  $D$  a matching degree between the value that  $D$  takes over  $x$  and the value that  $D$  takes over  $C$ . These matching degrees have been termed "marginal adequacy degrees". When the degrees are known, they are used by the system to compute the adequacy degree of object  $x$  to concept  $C$ . LAMDA [4] can model the concept of maximum entropy, that is, total homogeneity or indistinctness. This concept is represented by a class named Non-Informative-Class (abbreviated *NIC*). The adequacy of *NIC* is the same for all objects. The context determines the existence of this class and implies that *NIC* is always present in the space

of pre-established concepts. LAMDA considers the object that is being treated as an example of the concept with the highest adequacy degree. Then, it seems natural that in concept formation process, a class will be initialized any time. *NIC* is the concept with the highest adequacy degree to the object. Those observations such that their maximum adequacy degree is taken by *NIC*, are not assigned to any of the significative pre-established concepts in the recognition concept process. Then, *NIC* remains empty in the formation process, whereas *NIC* will be formed in the recognition process by those examples that cannot be described for the space of significative concepts, which is invariant in time. *NIC* plays the role of the minimum threshold of meaning traditionally used in classic clustering algorithms.

When an object is given to the system, the algorithm can be activated in two different modes. Recognition mode is used when concepts are already known. Otherwise, the algorithm makes a learning process that involves the formation of concepts. This learning process may be initialized by a set of pre-established concepts, or left without initial information, that is with *NIC* the only one initialized concept. Let us assume that when object  $x$  is given to the algorithm, the existing concepts are  $C_0, C_1, \dots, C_k$ , where  $C_0$  is *NIC*.

- Step 0 : For each component  $x_i$  of object  $x$ , get the extremum value  $x_{min}$  and  $x_{max}$  of the quantitative components. Compute the normalized value  $x_i$  for one element  $x$  for the descriptor  $i$  using the formula:

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

$i = 1 \dots n$ , where  $n$  is the number of attribute of  $x$  ;

- Step 1 : Compute the marginal adequacy degrees of  $x_i$  to all concept  $C_j$ , which is  $MAD_{x_i/C_j}$ ,  $j = 0, \dots, k$  et  $i = 1, \dots, n$  ;
- Step 2 : Compute the global adequacy (GAD) ;
- Step 3 : Search for the maximum computed degrees between all classes to deduce the possible owing class.

The fusion operation between the adequacy of concept is made using symmetrical sum:  $\alpha T + (1 - \alpha)C$  where  $T$  is the T-Norm and  $C$  is T-Conorm associated respectively with minimum and maximum operators.

Analysis of step 1. Two cases may occur :

1. Recognition mode. Object  $x$  is placed in  $C_j$ . If  $C_j$  is *NIC*, then  $x$  is said “unrecognized” ;

2. Learning mode. There are two possibilities:

Let  $\mu_j = \max_k(GAD(x/C_k))$

- $\mu_j$  does not correspond to *NIC*, in this case,  $x$  is placed in  $C_j$  and the representation of  $C_j$  will be modified to include  $x$ .

$$\mu_i = \mu_i + \frac{(x_i - \mu_i)}{N_{C_j} + 1} \quad (2)$$

where  $N_{C_j}$  is the number of elements in class  $C_j$  at this step of the algorithm.  $0 \leq j \leq k$

- $\mu_j$  corresponds to *NIC*. This means that  $x$  is the first element of a new class  $C_{k+1}$  and the representation of this new class will depend on  $x$ . We use the following formula to initialize the new class:

$$\mu_i = \mu_{NIC} + \frac{(x_i - \mu_{NIC})}{N_0 + 1} \quad (3)$$

where  $\mu_{NIC} = \mu_0$ .  $N_0$  is a fuzzy parameter and the larger it is, the greater number of elements will be accepted by  $C_j$ .

LAMDA treats objects sequentially. Each value descriptor contributes to the global adequacy of one object to one class through marginal adequacy degree (MAD). We use an fuzzy logic operator, which interpolate between union and intersection with an adjustable parameter called “exigency”, to add the descriptor information. Maximum exigency corresponds to the conjunction AND associates to intersection (The Minimum in this case). Let’s note that  $MAD = 1$  represents the total adequacy of one attribute to a class and  $MAD = 0$  represents the total inadequacy of

one attribute. Respectively, if the combination of MAD, which is the global adequation degree, is equal to 1, then the object has a total adequacy to the class.

By reference to the binomial probability in the binary case, the *MAD* of an object  $x$  such that  $w$  is the value of descriptor  $D$ , can be written as follow:

$$MAD(x/D) = \mu^w(1 - \mu)^{1-w} = f(\mu, w) \quad (4)$$

where  $\mu$  is the value that represents the class.  $w$  can be an object descriptor or a distance between the object descriptor and the corresponding class center.

The MAD function shown in equation (4) is needs only to the estimation of parameters which characterizes a class. These parameters can be estimated and modified when a new object is affected to the class.

## 2.1 T-LAMDA Clustering Algorithm

The notion of connections between the pixels is introduced by the notion of path and neighbor. A structure based on decision trees [5][6] is used to represent the classified data. This approach tries to classify a pixel according to some tests in each tree knot on the attributes which describe it. These tests are organized such as the response to one of them involves the next pixel test. The principle is to organize all the possible tests as a tree. A tree leaf indicates one of  $C$  classes (but to every class may correspond to several leaves) and to each node is associated a test (a selector) concerning an attribute, the R, G, B color values of pixels, their position or both.

In the data tree structure, only the father identifies the class and contains its parameters, the number of elements and the modality of its descriptors. Each node contains the parameters of the subclass he is the father. A forest is represented by a triplet weight (or size), father, modality ; “Size” will keep the number of parts of the partition, so we will be able to delete non significant region.

The Structure we used is like : Structure SET =

{ ValueR, ValueG, ValueB, SET \* Father, SET\* NextSet }.

The union cost is proportional to the height of the arborescence. Through “region growing” and tree representation method, fusions of regions are made. The “region growing” [7] is often used in segmentation of aerial images [8]. First, we start from small regions, which are either pixels or points and we group them until we consider that we are in the optimal case. Then, when a region becomes the son of another one it means that the root of the representative tree becomes the son of the element that identifies the father region. Region growing is often associated with union-find. Our operation find allows to go up to the tree father and identify the class. The Union allows to merge or not two elements. The Union operator uses the LAMDA fuzzy decision in equation 4.

In T-LAMDA algorithm, if the image size is  $N \times M$ , a number of  $N \times M$  classes are initialized at the beginning of the algorithm. These classes contain one pixel, each of the pixels in the image.

Our segmentation process makes fusion of two classes that correspond to two different regions, into one class. Let's take an example of 2 classes  $X$  and  $Y$  which parameters are  $\mu_X$  and  $\mu_Y$ . The number of pixels of each class is  $N_X$  and  $N_Y$ . We define the notion of class weight in relation to the number of pixels contained in this class. If  $N_X > N_Y$  the fusion of the classes will be made using the class  $X$  to compute the adequation degree. And if the next test  $GAD(\mu_Y|\mu_X) > GAD(\mu_0)$  is satisfied then the union will be made. If  $N_Y > N_X$  the fusion of the classes will be made using the class  $Y$  to compute the adequation degree. And if the next test  $GAD(\mu_X|\mu_Y) > GAD(\mu_0)$  is satisfied then the union of object  $X$  to class  $Y$  will be done.  $\mu_0$  is the parameter of the  $NIC$  and  $GAD(\mu_Y|\mu_X)$  represents the  $GAD$  of the class  $X$  in comparison with a class of weaker weight,  $Y$ . To construct the tree we use the update equation 2.

The characteristic algorithm is showed below:

- We sort the union pixels tests in order to test only adjacent regions, and to favor pix-

els whose attribute are close to each other. For example, let's take five objects  $A, B, C, D, F$ , see figure 1. Each point have the  $R, G, B$  feature. The neighbor objects are  $A, B, C, D$  according to figure 1. A sorted list containing the pair of connected elements according to the difference of the grey level  $\epsilon$  is made (figure 1) ;

- The distance selected for the presented result is the Maximum of the difference between each plane  $R, G, B$  :

$$\epsilon = \max(A_R - C_R, A_G - C_G, A_B - C_B)$$

The element connected, in the resulted tree, at the left in figure 1 are close spatially and have the same theoretical color. If  $A$  and  $B$  are close to each other, we take the decision relative to their own region. The construction of the region  $\{A, B, C, D\}$  is guided by the following plan, supposing that the possibility to merge  $A$  and  $C$  is stronger than the one to unite  $C$  and  $D$  and etc ;

- If the image size is  $N * M$ . The number of couples of pixels in all the image is  $(N * M)^2 - N * M$ . We use 4 and 8 neighborhood. In 4 neighborhood the number of pixels is  $(N - 1) * M + N * (M - 1)$ , and in 8 it is  $((N - 1) * M + N * (M - 1)) + (2 * (N - 1) * (M - 1))$ . In this method the union test with each pixel of the neighborhood is done supposing the first union test is superior to the last union test ;
- In addition to this, to obtain the final region of our image, the merging test are organized such as the possible union of the first couple of elements superior to the one of the last couple of pixels owing to the image ;
- At the end of the classification, we allocate arbitrarily a color at each class obtained from the mean of all the pixels belonging respectively to the region or a random color is affected to each region or class.

### 3 Biological Image Segmentation.

The images has been obtained using a Nikon video-camera coupled to a Olympus microscope

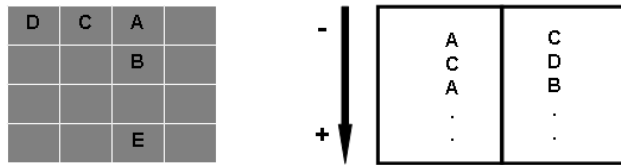


Figure 1: On the left we see the arrangement of the pixels in the plan. On the right, the sorted list. The smallest differences are placed towards the minus sign and the biggest differences are placed towards the plus sign.

X40. The cell's sample is the result of a biochemical reaction between a colorant and the cells. If we have a low metabolic activity, the coloring agent penetrates into the cells. It is going to change the original color in a dark blue color. So dead cells (symbolized D) are colored blue. In the case of microscopic images, we have the diffusion of the incident light on the cell's membrane. Of course, our clustering method needs to take this problem into account, which means to tune some parameters. We showed that only fuzzy decision tree give satisfactory results. It allows to diagnostic the microbial cell's evolution into a bioreactor. Based on the knowledge of the acquisition, we have defined a triangular membership function for the image background. We use a flooding algorithm associated with an edge detection to easily distinguish the background from the cells. The background is labelled with the fill-color. The flooding algorithm operates along the largest found area, which is the background. Different filter are used  $((Image - Filter * coeff) + Image)$  for flooding in order to obtain a buffered region with the cells localization. The flooding works in different ways : a queue algorithm, a recursive and a linear one. The results for a linear flooding are presented. The enhancement of the image allows to increase the gray level between image background and the analyzed region of the image.

*The flooding step using eight neighbourhood.*

```
Linear8(Im) {
p : starting point of Im, p(x,y)
1)FIND LEFT EDGE OF
COLOR AREA int LFillPts=x ;
//the location to check/fill on the left
```

```
ptr=p ;
//is the pointer to the current location
while(true)
{
ptr[0]=fillcolor ; //fill with the color
PixelsChecked[LFillPts,y]=true ;
LFillPts-- ; //decrease counter
ptr--1;//decrease pointer
if(LFillPts<=0) ||
(pixel is in the area defined
by a start color)||
(pixel have not been checked:
PixelsChecked[LFillPts,y]=false)
break ;
//exit loop if we're at edge
//of bitmap or color area
}
LFillLoc++ ;
1)FIND RIGHT EDGE OF COLOR AREA
2)START THE LOOP
UPWARDS AND DOWNWARDS
//By making the same thing
}
```

#### 4 Comparisons and Experiments

This new method has been compared with traditional algorithm : Mean-Shift, Fuzzy C-Mean, and Watershed.

Table 1: Data result obtained from the algorithm T-LAMDA. A : alive ; D : dead.

X data of the center	Y data of the center	Cell surface in pixels	Longest axis	State
99	79	1016	49	A
133	97	1744	47	A
202	96	756	54	D
226	114	1019	42	A

Figure 2(b) is obtained using FCM on color component. Applying the FCM algorithm on X, Y components (see figure 2(c)) allows us to separate connected cells, but the result is not often the optimum. We tried to solve the problem using a local perturbation on the center of the classes. The time of convergence increase, but it is a problem to go towards a real time application. The best

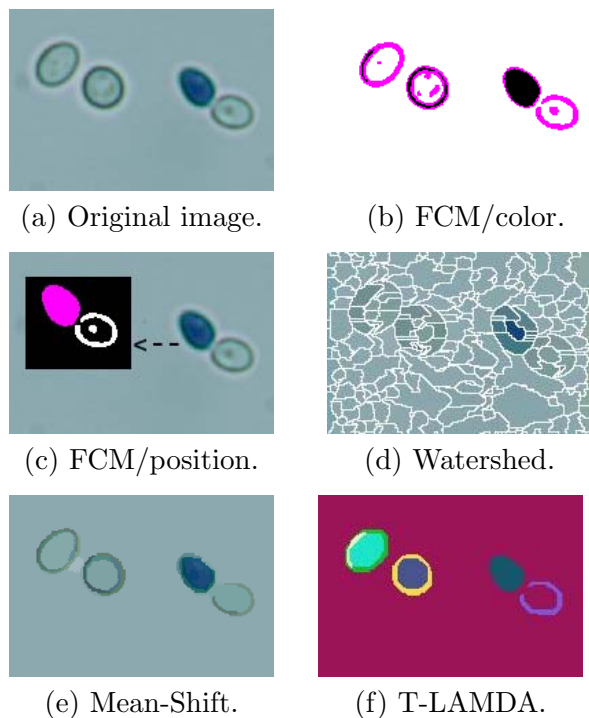


Figure 2: Result obtained with, watershed, mean-shift, fuzzy c-mean and T-LAMDA methods.



Figure 3: T-LAMDA result, using XY attributes. The results are presented with a random color per each class for a better observation.

results are obtained when each cells area is treated with a different FCM. Interesting results should be obtained with watershed (on figure 2(d)), but we have too many regions. A combined process could be used, but some cells could disappear due to cells regions apparently identical to adjacent region of the background. As Watershed, Mean-Shift on figure 2(e) give us too much regions, a quantity about 26. The Mean-Shift binary used is from Dorin Comaniciu et Peter Meer [9][10][11]. Watershed algorithm is very sensitive to noise therefore in our case, the bad results are due to the noisy images.

T-LAMDA algorithm produces 7 classes, but the diffusion problem remains. The method is less

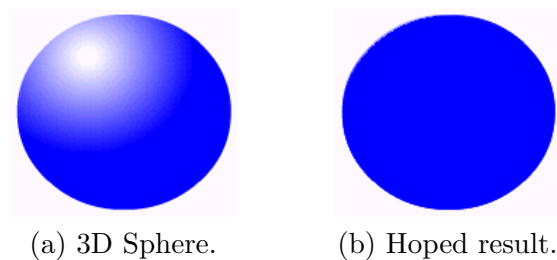


Figure 4: Reference images.

Table 2: Result obtained using different signal-noise ratio. Result presented are errors for T-LAMDA, Mean-Shift, Fuzzy C-mean, LAMDA.

Signal/ Noise in dB	T-LAMDA Error	Mean-Shift Error	FCM Error	LAMDA Error
11.33	0	0.67	0.5	0.90
12.61	0.5	0.67	0.5	0.91
10.21	0.5	0.75	0.57	0.92
11.05	0.5	0.88	0.6	0.92
10.10	0.5	0.93	0.69	0.95

sensitive to noise and smooth variation in the background.

The flooding steps allow us to improve the classification. We can overcome diffusion problem due to the light which gets through the cell towards the capture camera coupled to a microscope.

Figure 3 shows the classification result using T-LAMDA with X,Y component after the cells position were found. Four cells are identified. For a better observation, the results are presented with a random color for each class. Table 1 shows the result obtained with T-LAMDA using R, G, B, X, Y attributes.

We have done some tests using an image of a 3D sphere under the light. There is a strong diffraction of the light on the sphere ( see figure 4(a)). The goal of the study is to close the sphere to have in the best case the image in figure 4 (b).

Table 2 shows the result obtained for the different algorithms T-LAMDA, Mean-Shift, Fuzzy C-Mean, LAMDA. The Error term vary from 0 to 1 where 0 is the best result. The number of error of the FCM is caused by the fixed number of classes.

Two joined cells are separated by a priori knowledge of the cells position obtained after the segmentation process.

The algorithm is part of a .NET application named cell classification application (CELCA)<sup>1</sup>. The .NET technology is used in our application because it offers numerous libraries for software development. ADO.NET of .NET supplies a simple access in the data independently of the platform. XML becomes the standard format of transmission of data and any application capable to read the XML format can treat these data. More than 800 images are treated and the result is then presented in the XML form to exchange data with others applications.

## 5 Future Work and Improvement

MAD is computed in T-LAMDA algorithm using equation 4, we should compare the result using other function like Mahalanobis, Manhattan, or Euclidian distance or a Gaussian function. We should improve Watershed and Mean-Shift doing a hierarchical merging of the regions obtained after classification.

## 6 Conclusion

We present in this paper a clustering algorithm based on fuzzy region merging. This new method has been compared with traditional algorithm : Mean-Shift, Fuzzy C-Mean, and Watershed. The ability to work on the low level image definition has been proved. For example, biological systems analysis by microscope. These good results are due to the update of the memberships function when an instance is associate to a class and the possibility to follow a particular order in pixels merging by neighborhood. The low runtime allows us to use this software coupled to an in-situ microscope for the diagnostic of microbial populations.

<sup>1</sup>The service web is available at <http://perso.wanadoo.fr/atine.jc/>. The results are in the XML form.

## References

- [1] C. Olaru, L. Wehenkel, "A complete fuzzy decision tree technique," *Fuzzy Set and Systems* 138, pp. 221-254, 2003.
- [2] C. Z. Janikow, "Exemplar Learning in Fuzzy Decision Trees," Department of Mathematics and Computer Science, University of Missouri - St. Louis, 1996.
- [3] C. Marsala, "Construction d'arbres de décision flous : le système Slammbô," *Rencontre Francophone sur la logique floue et ses applications*, 1998.
- [4] J. Aguilar-Martin, F. Jarachi, M. Chan, "Partitioned identification techniques from poisson observations : application to cerebral blood flow estimation," Report LAAS No 89244, 11th IFAC World Congress, Tallinn (URSS), pp. 24-27, 13-17 August 1990.
- [5] "Apprentissage artificiel, concepts et algorithmes.," Ed. EYROLLES, pp. 334-362.
- [6] C. Fiorio, J. Gusted, "Two linear time Union-Find strategies for image processing.," LIRMM, october 1994.
- [7] M. LI, Ishwar K. Sethi, D. Li and N. Dimitrova, "Region Growing Using Online Learning.," *CISST* pp. 73-76, 2003.
- [8] M. Bicego, S. Dalfini, V. Murino, "Extraction of geographical entities from aerial images," *Proc. of IEEE-ISPRS Workshop on Remote Sensing and Data fusion over Urban Areas URBAN '03*, Berlin, Germany, 125-128, June 2003.
- [9] D. Comaniciu, P. Meer, "Mean Shift : A Robust Approach toward Feature Space Analysis," *IEEE-PAMI*, vol. 24, 603619, 2002.
- [10] D. Comaniciu, P. Meer, "Mean Shift Analysis and Applications," *Proc. of ICCV*, pp. 1197-1203, 1999.
- [11] H. Wang and D. Suter, "Color Image Segmentation Using Global Information and Local Homogeneity," *Proc. VIIth Digital Image Computing : Techniques and Applications*, Sydney, 10-12 Dec. 2003.