Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Unsupervised Data Clustering and Image Segmentation using Natural Computing Techniques

Jackson G. de Souza, José Alfredo F. Costa
Adaptive Systems Laboratory, Department of Electrical Engineering, Center for Technology
Federal University of Rio Grande do Norte
Natal - RN, Brazil
jgomes@ulbra-to.br, alfredo@dee.ufrn.br

***Abstract*— Natural computing (NC) is a novel approach to solve real life problems inspired in the life itself. A diversity of algorithms had been proposed such as Evolutionary Techniques, Genetic Algorithms and Particle Swarm Optimization (PSO). These approaches, together with fuzzy and neural networks, give powerful tools for researchers in a diversity of problems of optimization, classification, data analysis and clustering. This paper presents concepts and experimental results of approaches to data clustering and image segmentation using NC approaches. The main focus are on Evolutionary Computing, which is based on the concepts of the evolutionary biology and individual-to-population adaptation, and Swarm Intelligence, which is inspired in the behavior of individuals, together, try to achieve better results for a complex optimization problem. Genetic and PSO based K-means and fuzzy K-means algorithms are described. Results are shown for data clustering using UCI datasets such as Ruspini, Iris and Wine and for image texture and intensity segmentation using images from BrainWeb system.**

***Keywords*—unsupervised data clustering, image segmentation, evolutionary techniques, genetic algorithms, particle swarm optimization, natural computing.**

I. Introduction

Natural computing (NC) is a novel approach to solve real life problems inspired in the life itself. A diversity of algorithms had been proposed such as evolutionary techniques, genetic algorithms and particle swarm optimization (PSO). These approaches, together with fuzzy and neural networks, give powerful tools for researchers in a diversity of problems of optimization, classification, data analysis and clustering.

Clustering methods are usually stated as methods for finding the hidden structure of data. A partition of a set of *N* patterns in a *p*-dimensional feature space must be found in a way that those patterns in a given cluster are more similar to each other than the rest. Applications to clustering algorithms range from engineering to biology [1-3].

Image segmentation techniques are based on Pattern Recognition concepts and such a task aims to identify behavior in a data set. In the context of image segmentation, the data set represents image data, coded as follows: the light intensity value (the pixel data) represents a pattern, an item in the data set, and the color information is represented by columns (the feature vectors). Clustering techniques represent the non-supervised pattern classification in groups [3]. Considering the image context, the clusters correspond to some semantic meaning in the image, which is, objects. More than simple image characteristics, these grouped semantic regions represent information; and image segmentation is applicable in an endless list of areas and applications, for example: computer-aided diagnosis (CAD) being used in the detection of breast cancer on mammograms [4], outdoor object recognition, robot vision, content-based image, and marketplace decision support.

Among the many methods for data analysis through clustering and unsupervised image segmentation is: Nearest Neighbor Clustering, Fuzzy Clustering, and Artificial Neural Networks for Clustering [3]. Such bio and social-inspired methods try to solve the related problems using knowledge found in the way nature solves problems. Social inspired approaches intend to solve problems considering that an initial and previously defined weak solution can lead the whole population to find a better or a best so far solution.

This paper presents concepts and experimental results of approaches to data clustering and image segmentation using (NC) approaches. The main focus are on Evolutionary Computing, which is based on the concepts of the evolutionary biology and individual-to-population adaptation, and Swarm Intelligence, which is inspired in the behavior of individuals, together, try to achieve better results for a complex optimization problem. Genetic and PSO based K-means and fuzzy K-means algorithms are described. Results are shown for data clustering using UCI datasets such as Ruspini, Iris and Wine and for image texture and intensity segmentation using images from BrainWeb system.

The remainder of the paper is organized in the following form: section 2 describes briefly data clustering and image segmentation; section 3 approaches natural computing and section 4 focuses clustering using natural computing methods. Section 5 presents experimental results and discussion and section 6 gives the conclusions and final considerations.

II. Data Clustering and Image Segmentation

To find clusters in a data set is to find relations amongst unlabeled data. The "relation" means that some data are in some way next to another that they can be grouped. It is found in [3] that the components of a clustering task are:

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

1. Pattern representation includes: feature selection, which identifies the most effective subset of the original features to use in clustering; and feature extraction, which is the preprocessing of the input features.

2. A Distance measure is used to determine pattern proximity. A simple, and, perhaps, the most used, distance function is the Euclidean Distance.

3. Clustering relates to finding the groups (or, labeling the data) and it can be hard (an element belongs to one group only) or fuzzy (an element belongs to one group following a degree of membership).

4. Data abstraction is an optional phase and extracts a simple and compact representation of a data set and, in the case of data clustering, some very representative patterns are chosen: the centroids.

5. Assessment of output is the process of evaluating the clustering result. Cluster validation techniques are, also, a traditional approach to dynamic clustering [5].

III. Natural Computing

Natural computing "is the computational version of the process of extracting ideas from nature to develop computational systems, or using natural materials to perform computation" [6] and some most representative approaches are the following:

Artificial Neural Networks. An artificial neural network, as found in [7], is a massively distributed parallel built-in processor composed of simple processing units (the neurons) that act, naturally, to store useful knowledge which is acquired through a learning process that yields better results when the processing units work in a network interconnected form (the neural network).

Evolutionary Computing. The ideas of evolutionary biology and how descendants carry on knowledge from their parents to be adaptive and better survive are the main inspiration to develop search and optimization techniques for solving complex problems.

Swarm Intelligence. As an example technique "Particle Swarm Optimizers (PSO) are population-based optimization algorithms modeled after the simulation of social behavior of bird flocks" [5].

Artificial Immune Systems. This technique refers to adaptive systems inspired by theoretical and experimental immunology with the goal of solving problems.

IV. Clustering using Natural Computing

This section presents two methods based on this concept: the Genetic K-Means Algorithm and Particle Swarm Optimization both used in data clustering and image segmentation.

A. Genetic K-Means Algorithm

Genetic Algorithms have been applied to many function optimization problems and are shown to be good in finding optimal and near optimal solutions [8]. Aiming to solve the partitional clustering algorithm problem of finding a partition in a given data, with a number of centroids (or clusters), Genetic K-Means (GKA) is introduced by [8]; it establishes an evaluation criterion based on the minimization of the Total Within Cluster Variation (TWCV), an objective function that is defined as follows [4], [9].

Given **X**, the set of N patterns, and X_{nd} the dth feature of a pattern Xn, G_k the kth cluster and Z_k the number of patterns in G_k, the TWCV is defined as:

$$TWCV = \sum_{n=1}^{N}\sum_{d=1}^{D} X_{nd}{}^{2} - \sum_{k=1}^{K}\frac{1}{Z_k}\sum_{d=1}^{D} SF_{kd}{}^{2} \qquad (1)$$

where SF_{kd} is the sum of the dth features of all patterns in G_k. The TWCV is also known as *square-error measure* [8]. The objective function, thus, tries to minimize the TWCV, finding the clustering that has centroids attending concepts of [5] *compactness* (patterns from on cluster are similar to each other and different from patterns in other clusters) and *separation* (the clusters' centroids are well-separated, considering a distance measure as the Euclidean Distance). It is found in [10] another method for genetic algorithm based clustering that uses another fitness function, the Davies-Boudin index, which is a function of the ration of the sum of within-cluster scatter to between-cluster separation. As will be seen later, other validation indexes may be used and despites the objective function, GKA main aspects are:

1. *Coding*. Refers to how to encode the solution (the chromosome); one way of doing this is the *string-of-group-numbers encoding* where for Z coded solutions (partitions), represented by strings of length N, each element of each string (an allele) contains a cluster number.
2. *Initialization*. The initial population P_0 is defined randomly: each allele is initialized to a cluster number. The next population P_{i+1} is defined in terms of the selection, mutation and the K-means operator.
3. *Selection*. Chromosomes from a previous population are chosen randomly according to a distribution.
4. *Mutation*. The mutation operator changes an allele value depending on the distances of the cluster centroids from the corresponding pattern.
5. *K-Means Operator* (KMO). This operator is used to speed up the convergence process and is related to one step of the classical K-means algorithm. Given a chromosome, each allele is replaced in order to be closer to its centroid.

Another approach, K-Means Genetic Algorithm (KGA), is presented in [10] and shows a slight modification to the definitions presented before: the crossover operator is added to the algorithm and it is a probabilistic process that exchanges information between two parent chromosomes for generating two new (descendant) chromosomes.

B. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a population based stochastic algorithm modeled from the observation and simulation of bird flocks behavior [11]. The difference to evolutionary computation is that, in PSO, each particle benefits from its own previous solutions (historic) and there is no such approach in evolutionary methods [11].

The model of adaptive culture and particle swarms that drives PSO is based on three main principles [11]: to evaluate, to compare, and to imitate.

The classical definition is that each particle, amongst the multitude of individuals (the swarm), flies through the search space [11] and carries on a potential solution to the optimization problem [5]. The movement of particle, i.e. the changing of position, is determined by an equation that considers the current position of the particle and a velocity vector [5], [11]:

$$\mathbf{x}_i = \mathbf{x}_i(t) + \mathbf{v}_i(t+1) \tag{2}$$

$$\begin{aligned}\mathbf{v}_i(t+1) = \varpi \mathbf{v}_i(t) + c1r1\left(\mathbf{p}_i(t) - \mathbf{x}_i(t)\right) + \\ c2r2\left(\mathbf{p}_g(t) - \mathbf{x}_i(t)\right)\end{aligned} \tag{3}$$

where, according to [5], ϖ is the inertia weight, which controls the impact of the previous velocity; c1 and c2 are acceleration constants and r1~U(0,1) and r2~U(0,1); $\mathbf{p}_i(t)$ is the best position of particle *i*; and $\mathbf{p}_g(t)$ is the best position among all particles of the swarm. A user defined maximum velocity can be used to constraint the velocity update. The performance of the particle is measured using a fitness function which depends on the optimization problem.

C. Clustering using Particle Swarms

Different approaches are found that implement clustering based PSO algorithms, such as [5] and [11]. A PSO-based Clustering Algorithm (PSOCA) can be defined as follows [5], [11]: in the context of data clustering, a single particle represents de set of *K* cluster centroids, in other words, each particle represents a solution to the clustering problem and, thus, a swarm represents a set of candidate data clusterings. The main steps are: a) initialize particle position and velocity (for each particle); b) while a stop criterion is not found, for each particle: calculate particle quality; find particle's best and global best; and update the velocity of particle.

V. Experimental Results and Discussions

The experiments rely on evaluate numerical results of clustering algorithms based on Genetic Algorithms and PSO. As previously seen, both methods are modeled to allow a switch of the traditional and basic clustering algorithm. Thus, this allows us to define the following algorithms variations:

a) Genetic K-means [Clustering] Algorithm (GKA);

b) Genetic Fuzzy C-means [Clustering] Algorithm (GFCMA);

c) PSO-based K-means [Clustering] Algorithm (PSOKA); and

d) PSO-based Fuzzy C-means [Clustering] Algorithm (PSOFCMA).

The datasets used in data clustering experiments are the following:

a) Ruspini: two-dimensional dataset with 75 patterns; has four classes easily separable;

b) Wine: thirteen dimensions and 178 patterns; has three classes; and

c) Iris: four-dimensional dataset with 150 patterns; has three classes.

To best evaluate the results, considering classification error, in each dataset was added another dimension, corresponding to the cluster number associated to the pattern. Cluster validation indexes were used to obtain numerical results and guide the possible best solution found by the algorithms: Davies-Bouldin (DB), SC, S, and Xie-Beni (XB). Although data clustering were realized first to guide and support image segmentation experiments, their results will not be presented here for limited space and paper focus reasons. Fig. 1 shows data clustering results for Wine dataset using GFCMA method (using PCA to reduce dimensions).

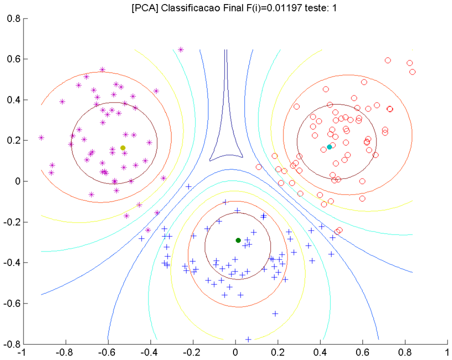


Figure 1. GFCMA results for Wine dataset.

The dataset used in image segmentation experiments was obtained from the BrainWeb system [12], it corresponds to simulated MR images of T1 modality, 0% noise, and 0% intensity. BrainWeb dataset contains 10 classes that range from background to connective material. For ground truth and classification error evaluation is used the "fuzzy" dataset. Beyond the original dataset (called image_intensity), to evaluate the behavior of the algorithm on texture information, another dataset was created, called image_texture, which adds 3 new dimensions corresponding to mean, standard deviation and variance considering a 5x5 window.

For comparison purposes experiments were taken for classical K-means and Fuzzy C-means (FCM) algorithms. Results are presented by Table 1. The values correspond to classification (image segmentation) error.

TABLE I. Image Segmentation Results for K-means and FCM

Dataset	K-means	FCM
Image_texture	62,0263± 6,3218	69,3057± 2,1603
Image_intensity	67,3747± 6,3854	70,0379± 3,4461

Numerical results for image segmentation are shown from Table 2 to Table 9, considering GKA, GFCMA, PSOKA, and PSOFCMA.

TABLE II. GKA IMAGE SEGMENTATION ON IMAGE_TEXTURE

CVI	**GKA**		
	Min.	***Mean.***	***Max.***
DB	0,31476	0,33616	0,35755
SC	0,16075	0,16075	0,16075
S	0,00001	0,00001	0,00001
XB	169,59	202,84	236,09
Error (%)	75,19	85,17	90,80

TABLE III. GFCMA IMAGE SEGMENTATION ON IMAGE_TEXTURE

CVI	**GFCMA**		
	Min.	***Mean.***	***Max.***
DB	0,33990	0,35202	0,36414
SC	0,16075	0,28419	0,29732
S	0,00001	0,00001	0,00001
XB	84,10	92,94	101,78
Error (%)	55,66	61,42	70,80

TABLE IV. GKA IMAGE SEGMENTATION ON IMAGE_INTENSITY

CVI	**GKA**		
	Min.	***Mean.***	***Max.***
DB	0,68072	0,68072	0,68072
SC	0,20654	0,20654	0,20654
S	0,00001	0,00001	0,00001
XB	10,19488	10,19488	10,19488
Error (%)	84,96	87,68	91,68

TABLE V. GFCMA IMAGE SEGMENTATION ON IMAGE_INTENSITY

CVI	**GFCMA**		
	Min.	***Mean.***	***Max.***
DB	0,71655	0,73323	0,74991
SC	0,64346	0,64346	0,64346
S	0,00003	0,00003	0,00003
XB	1,17901	1,24838	1,31775
Error (%)	67,67	69,49	71,27

TABLE VI. PSOKA IMAGE SEGMENTATION ON IMAGE_TEXTURE

CVI	**PSOKA**		
	Min.	***Mean.***	***Max.***
DB	0,56977	0,63134	0,69290
SC	0,25217	0,25217	0,25217
S	0,00001	0,00001	0,00001
XB	1,85475	1,85475	1,85475
Error (%)	45,18	72,28	90,06

TABLE VII. PSOFCMA IMAGE SEGMENTATION ON IMAGE_TEXTURE

CVI	**PSOFCMA**		
	Min.	***Mean.***	***Max.***
DB	0,67515	0,69406	0,71297
SC	0,54645	0,55021	0,55397
S	0,00002	0,00002	0,00002
XB	1,03556	1,08944	1,14332
Error (%)	57,78	66,60	72,69

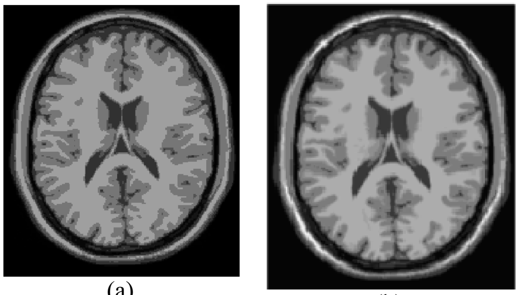
TABLE VIII. PSOKA IMAGE SEGMENTATION ON IMAGE_INTENSITY

CVI	**PSOKA**		
	Min.	***Mean.***	***Max.***
DB	0,29822	0,32084	0,34345
SC	0,12357	0,15516	0,18674
S	0,00001	0,00001	0,00001
XB	210,67	235,639	260,6045
Error (%)	60,57	76,13	88,45

TABLE IX. PSOFCMA IMAGE SEGMENTATION ON IMAGE_INTENSITY

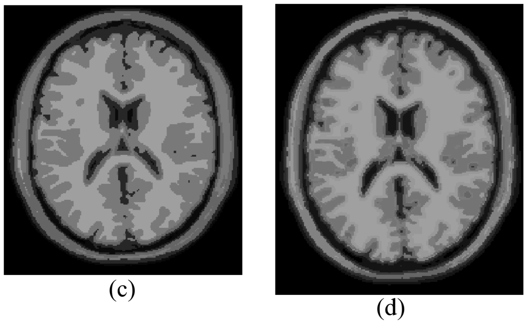
CVI	**PSOFCMA**		
	Min.	***Mean.***	***Max.***
DB	0,34237	0,34627	0,35017
SC	0,12357	0,27962	0,29527
S	0,00001	0,00001	0,00001
XB	81,4605	83,4044	85,34844
Error (%)	65,87	70,00	72,86

Qualitative image segmentation results for GKA, GFCMA, PSOKA, and PSOFCMA are presented by Fig. 2 and Fig. 3.



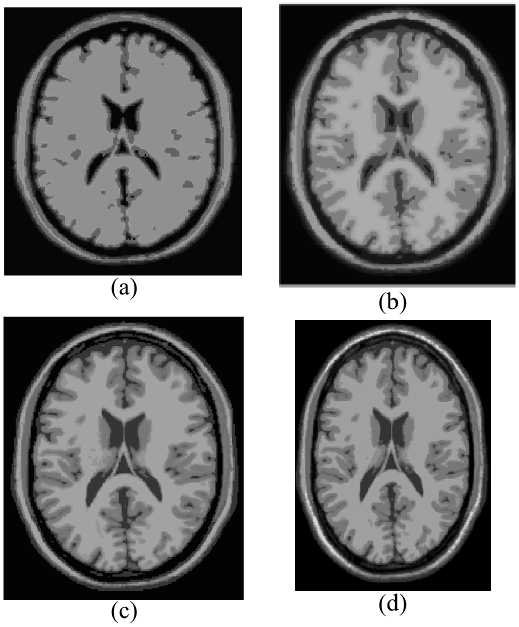
(a)

(b)



(c) (d)

Figure 2. Image segmentation results for image_texture and image_intensity: GKA (a, c); GFCMA (b, d).



(a) (b)

(c) (d)

Figure 3. Image segmentation results for image_texture and image_intensity: PSOKA (a, c); PSOFCMA (b, d).

VI. CONCLUSIONS

The present paper presents two natural computing methods for data clustering and image segmentation, their implementation and some results, one based on Genetic Algorithms and the other based on Particle Swarm Optimization. The task of image segmentation is not a trivial process. Considering the medical imaging context it is highly important the specialist's opinion about the results found. As the dataset is simulated the experiments were guided by this situation. Thus, it is necessary to make experiments with real imagery.

The methodology used in this paper was based on the following: 1) to implement the algorithms; 2) to evaluate clustering results on known databases; 3) use the obtained results to guide tests with image segmentation. Image segmentation tests must consider image characteristics. Thus two datasets were created: one that considers image data as is; and another that considers texture information. As the present methods are based on Evolutionary Computation and all have a performance (fitness) function, there must be some way to guide this evolution, so tests were made considering several Clustering Validation Indexes: DB, SC, S and XB. Also, a measure of classification error was used to identify the method's final and overall performance. CVI can be used as a function of quality of a solution (population/generation for GKA or particle for PSO).

Quantitative analysis shows that: a) considering CVI and image_texture dataset, GKA got best results considering indexes DB, SC, and S, while PSOFCM got best results for index XB; b) considering CVI and image_intensity dataset, PSOKA got best results for indexes DB, SC and S, while GFCMA got best results for index XB; c) considering Classification Error GFCMA got best results for image_texture and image_intensity datasets.

Qualitative analysis must take in account that the image data is not real, but simulated (as stated before), so, it can be seen that GKA and PSOFCMA got best results for image_texture and image_intensity datasets. To evaluate this, the classified element's region is take into account, and how much they are likely the related regions in original image. Although an in depth analysis with specialist's support is also necessary to enforce such argumentation.

Another discussion is about the values obtained for classical K-means and FCM algorithms. Numerical results show the inferiority of K-means in the context of GA and PSO, in contrast to values obtained by classical K-means. In counterpart, values obtained using FCM (for GA and PSO) are a little bit lower than the ones obtained by classical FCM, but not so much to state GA and PSO overall superiority. Most experiments using classical K-means and FCM run to 100 iterations – and more iteration could lead to lower error values. It's necessary to remember that GA and PSO both use only one iteration of K-means and FCM, and the convergence is fast (about 5 to 10 iterations). The problem of possible premature convergence of PSO is investigated by [13], which proposed the Improved PSO (IPSO) algorithm. This is a problem to take into account as a try to improve image segmentation results for PSO and GA also.

In summary, considering the results obtained from the tests, it can be said that methods based on FCM performed better considering Classification Error, and K-means was better considering CVI. As the present work does not evolves to image registration and classification more evaluation is necessary to argue about Fuzzy C-means superiority over K-means, in terms of the implemented algorithms.

Acknowledgment: This research was in part funded by the Brazilian Science Council (CNPq), grants 480043/2008-6 and 201382/2008-3.

REFERENCES

[1] Xu, R. and Wunsch II, D. "Survey of Clustering Algorithms", *IEEE Trans. on Neural Networks*, Vol.16, Iss.3, pp. 645-678, May 2005.

[2] Xu, R. and Wunsch II, D. *Clustering*. Wiley, 2008.

[3] Jain, A. K., Murty, M. N., Flynn, P. J.. Data clustering: a review. ACM Computer Surveys (31), ACM Press, 264--323 (1999)

[4] Doi, K.. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. Comp. Medical Imaging and Graphics, vol. 31, Elsevier, 198--211 (2007)

[5] Omram, M. G., Salman, A., Engelbrecht, A. P.. Dynamic clustering using particle swarm optimization with application in image segmentation. Pattern Anal. Applic. (8), Springer-Verlag, London, 332--344 (2006)

[6] Castro, L. N. de. Fundamentals of natural computing: an overview. Physics of Life Rev. 4, Elsevier, 1--36 (2007)

[7] Haykin, S.. Neural Networks: a comprehensive foundation, 2nd ed. Prentice Hall (1998)

[8] Krishna, K., Murty, M. N.. Genetic K-Means Algorithm. IEEE Trans. Systems, Man, and Cybernetics – Part B: Cybernetics, vol. 29, no. 3, 433--439 (1999)

[9] Lu, Y., Lu, S., Fotouhi, F., Deng, Y., Brown, S. J.. Incremental genetic K-means algorithm and its application in gene expression data analysis. BMC Bioinformatics, BioMed Central (2004)

[10] Bandyopadhyay, S., Maulik, U.. Genetic clustering for automatic evolution of clusters and application to image classification. Pat. Recog (35), Elsevier, 1197--1208 (2002)

[11] Omram, M. G. H.. Particle Swarm Optimization Methods for Pattern Recognition and Image Processing, Ph.D. Thesis, University of Pretoria, Pretoria (2004)

[12] BrainWeb: Simulated Brain Database. Available online at: <http://www.bic.mni.mcgill.ca/brainweb/>.

[13] Yong-gang, L., Wei-hua, G., Chun-hua, Y., Jie, L.. Improved PSO algorithm and its application. Journal of Central South University of Technology, vol. 12, number 1, 222—226 (2005)