AN EFFICIENT ANT COLONY OPTIMIZATION CLUSTERING ALGORITHM

Fadi ABDIN, Angel OSORIO
LIMSI-CNRS, B.P. 133
91403 Orsay CEDEX, FRANCE
abdin@limsi.fr , osorio@limsi.fr

ABSTRACT

This paper presents a new algorithm for clustering which is called an "efficient ant colony optimization clustering algorithm" (EACOC**)** based on a classic algorithm "LF algorithm". We have proved the algorithm efficiency when dealt with a big variety of different data as well as providing high quality and converging speed simultaneously. This is considered as the outcome of many changes we have made including redefining the digital manner of ants, setting new formula to find out the degree of similarity and measuring the distance between objects; as well as creating a process to determine the degree of similarity between the collections resulting from the repeated processes. Experimental results show, by using clustering benchmarks indicate, that this suggested algorithm is the best of (LF) Algorithm, as it could defeat the defects found in (LF) involving; the law converging speed and the big number of repeating processes.

Keywords: Ant colony algorithm, BM algorithm, LF algorithm; iris data.

1 INTRODUCTION

The ants behavior is one of the most typical samples of the intelligent flocks. The ant is a relatively simple creature .It is unable to do complicated missions; however, the hard work of the whole ants colony has the ability to perform much better than individual performances [1]. Through examining the behavior within the ants' colonies, and its capacity of self-organizing and exchanging data (information), we can say that the ants' colony can do complicated jobs, which go beyond the individual ability of a single ant. We have discovered a number of vital properties such as flexibility, solidity, self-organizing, and non-centrality. These latter are widely applicable on solving the optimization problems, communication networks, data mining, etc [2-3].

The process of data mining can be defined as the process of finding out right meaningful information involved in the collection of raw data. It is a strong, powerful tool used to automatically divert a huge amount of data into useful and common information. Besides, analyzing data collections plays a great part in data mining applications, as well as it is considered as one of the most important researches in this domain. It also determines the properties included in the data collections as each one of these collections consists in the same objectives between each other, and different from the objectives of other groups [4-6].

The Algorithm of ants colony is a new digital algorithm depending on the smart digital flocks which are inspired by the organized behavior of the insects which live an interactive life style, since the real ants are known for two distinct styles : the foraging and clustering. Both researches Dorigo & Al [7] could find out, through the former conduct, an algorithm to search for the perfect route in the colony which is called Ant Colony Optimization (ACO). It has been used to be done with the problems of discrete examples. According to the clustering conduct which was noticed through heaping bodies (corpuses) of ants , that the first inspired work was the clustering algorithm of the ants colony (BM) by the searchers Al & Deneubourg [8]. This latter doesn't need an ex-quittance of the number of groups (collections) through which the clustering process will be done. It provides, through using the arbitrary search mechanism, a good converging speed and doing a complete and perfect global search. Moreover, the clustering conduct of ants' colony is close to the process of analyzing data collections. Therefore, this algorithm has got a normal characteristic and can be mainly applied to be done with data collecting issues.

Depending on the main model (BM), both researchers Lumer & Faieta have worked to find a formula to measure similarity between two objects of data collections [9]. Thus, (BM & LF) became very common samples used on many scales [10-13]. Due to the huge number of parameters which must be

determined previously, as well as some useless arbitrary movements, doing these models require a long time to perform and waste a big space of memory. To avoid these obstacles, we have offered our algorithm (EACOC) in this paper.
We have made a new form for the digital conduct of ants and to measure the similarity degree. We have also set a new method to suit the collections which are considered as an outcome from every repeating process, experimental results show that the suggested algorithm can reduce the number of repeating times, and then the duration of performance.
Through this paper we will narrate the following; section 2: We are going to show the bonds between the complete work and reality. Section 3: we will present the modifications that we have executed in our algorithm. Section 4: We are going to show the practical consequences, and make a comparison with the same consequences of the same collection of data applied in (LF) Algorithm. Section 5: briefly shows the conclusion and the future works.

2 HEAPING BODIES IN THE ANTS' COLONY AND ITS PRACTICAL APPLICATIONS

Several worker ants, for many species of ants, have the task of piling bodies and sorting the lavers, to clean the nests. We can mention the experiment performed by Chrétien [14] on the Lasiusniger Ants, as an example of ants' types that organize graves. We can also provide the study done by Al & Deneubourg on the following types: Pheidole pallidula and Messor sancta concerning the construction of cemetery.

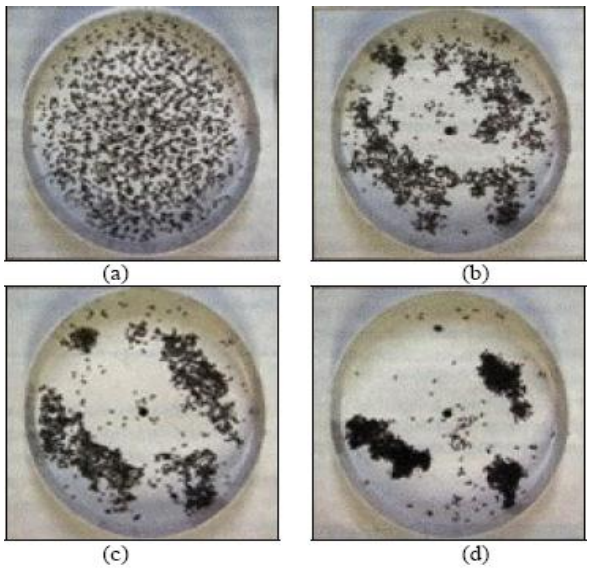

Fig.1: a sequential clustering task

Figure (1) shows, from (a) to (d) the process of heaping bodies in an experiment ants colony, as there are 1500 bodies spread chaotically on a circular space with a radius of (25cm). Messor Sancta workers ants perform this task, as we can see the initial case (a), (b) two hours later, (c) six hours later, and (d) twenty six hours after the beginning of the experiment. Through this experiment, it was noticed that if the experimental space isn't big enough, or even has got variable areas, these collections will spread on the edges of this area. It's going to be precisely identical with the variable areas. The main mechanism that we may bring out of this experiment is the process of approaching followed worker ants while clustering the corpuses collections, as it is done according to the different size through using the positive feeding. We can notice that each group grows more and more older due to the ants' environment which is known for (stigmergic) characteristic. Depending on this experimental research, Al & Deneubourg have suggested the vital clustering model in the ants' colony which is called (BM). The fundamental point of this model is that the digital ants lead an arbitrary movement in the form of a square net of cells, on which data collection are spread, without ignoring that a single cell can't have more than one object of data, as well as the digital ants can move on this net picking up or dropping these objects from a cell to another according to with the density of similar objects with the neighbor ones. The probability of picking up objects *Pp* for one free ant is given through the following equation:

$$P_p = \left(\frac{k_1}{k_1 + f} \right)^2 \qquad (1)$$

Where *f* is the function that we use to calculate the degree of similarity between the object and its neighbor and this occurs within the region that ants can perceive. *k1* is a probability constant that adjusts the ability of ants to picking up objects. When $f << k1$, *Pp* is close to 1,so that, the probability of picking up an item is high when there are not many similar items in the neighborhood, whereas, when $f >> k1$, *Pp* is close to 0, so that, the probability of picking up an item is low when there are many similar items in the neighborhood .
The probability of dropping objects *Pd* for a randomly moving loaded agent to deposit an item is given by:

$$p_d = \left(\frac{f}{k_2 + f} \right)^2 \qquad (2)$$

Where *k2* is another probability constant: when $f <<$

k2, *Pd* is close to 0,so that, the probability of dropping an item low when there are not many similar items in the neighborhood, whereas when $f >> k2$, *Pd* is close to 1,so that, the probability of dropping is high when there are many similar items in the neighborhood. f is found as shown below :

$$f = \frac{N}{T} \qquad (3)$$

As N represents the number of objects available nearby for every operation (process), and T represents the number of times kept by each one of the digital ants during every turn.
They have shown, through modeling, a production of small collections during a relatively short period of time which integrated to generate the final collections within a relatively long period of time. Then, objects are a lot like the one shown in figure (1)

Through generalizing the (BM) sample, Lumer & Faieta could set their model (LF), which is applicable on data analysis. The main point was to define the distance between the objects within a collection of data; in (BM) that's originally binary. In other words, if we have two objects *Oi*, *Oj*, which can be similar or different, then the distance between them can determine either if they are alike (0) or not (1). This approach can be widened to include more complicated data collections, so each object may has variance properties. Thus, finding out the distance is more sophisticated as we can use Euclidean Rule we consider the objects as a group of points in the Rn area. Therefore, if we assume that data collection of (LF) consists of (n) objects, each is represented by an array, all the objects will be spread on a square net (m*m), but there must be one component at the most in each cell of the net, as we consider that (m) value highly affects the performance of clustering process. If (m) is very big, then the objects will scatter on the net and the digital ants will be moving until they are able to hold or pick up an object, and then the duration of switching will increase every time, and the accurateness of similarity between the object and the neighbor ones will decrease. Moreover the number of digital worker ants will also affect the performance of clustering process. Then, if the number is huge, there will be too many digital holder worker ants and so, a lot of objects featured as (busy), i.e. these objects will be ignored when finding out the similarity degree of an object. Thus it will not only influence the accurateness of similarity degree, but will also cause laziness in clustering process, because worker ants won't find the proper place to put carried objects. Then they will have to carry these objects for longer time. On the other side, if the number is too small, then the converging speed of the algorithm will be reduced.
The similarity degree, when the digital ants move in an arbitrary manner on two dimensional spaces is given according to the following equation:

$$f(o_i) = \begin{cases} \frac{1}{\gamma^2}\sum\left(1 - \frac{d(o_i, o_j)}{\alpha}\right) & f(o_i) > 0 \\ 0 & otherwise \end{cases} \qquad (4)$$

When $f(Oi)$ is the similarity degree between *Oi* and its neighbors within a space of radius γ, which can be also called the space of vision for the digital worker ants, when γ is big , it needs longer time to tread . $d(Oi,Oj)$ represents the distance between *Oi* and *Oj* which is given cording the Euclidean method to find out the distance (space) in Mahalanobis [15] model, that follows :

$$d(o_i, o_j) = \sqrt{\sum (o_j - o_i)^2} \qquad (5)$$

α is the similarity coefficient which adapts the similarity average of the collection. α value will highly affects the number of groups and then the converging speed of algorithm. So, if α is very small, then two similar objects will be separated into two different groups, which will provide more number of groups, and law converging speed. However, it will be considered as a high quality clustering for one collection, but not appropriated in a global approach. On the contrary, if α is too big, then two distinct objects separated by big Euclidean distance will be collected in the same group. This may lead to less number of groups and higher converging speed, but low clustering quality. In short terms, selecting a proper value of α may make the space between objects in one group less than the space between different groups. To complete what we have just started we can define the probability conversion function as the used one to convert the average similarity to the probability value, this will enable the ants to decide whether to pick up or drop the objects, without ignoring that the probability conversion function must undergo the following conditions:

1- The similarity degree must be one of the parameters.
2- The relation between the picking up probability (*Pp*) and similarity degree is an inverse relationship, and between the dropping probability (*Pd*) and similarity degree is a direct relationship as it's shown in the following equations:

$$p_p = \left(\frac{k_p}{k_p + f(o_i)} \right)^2 \qquad (6)$$

$$p_d = \begin{cases} 2f(o_i) & f(o_i) \le k_d \\ 1 & otherwise \end{cases} \qquad (7)$$

Kp is the probability picking up constant that adapts the efficiency of a free digital worker ant to carry an object from its location. *Kd* is the probability dropping constant that adapts the efficiency of a holder digital worker ant to put the carried object at the placed it's moved to through an arbitrary movement. Figure (2) shows the mechanism through which the algorithm (LF) works, it demonstrates that when a free ant moves to a cell that includes an object, it will find out the value of picking up probability in accordance with (6). Then it will be compared to an arbitrary value to make a decision. On the other hand, if the ant is holder like it will find out the value of dropping probability when moving to an empty cell, according to (7), then compared with a random probability value.

/* Initialization phase*/
For every item *Oi* do
Place *Oi* randomly on grid
End For
For all ants do
Place ant at randomly selected site
End For
/* main loop */
For t=1 to *t*max do
For all ants do
If ((ant unladed) and (site *Sj* occupied by item *Oi*))then
Compute the similarity of *f(Oi)* in radius γ
Calculate the *Pp* and generate a random number Q
If (*Pp* > Q) /*pick-up rule*/
Pick up item
Move the ant with the item *Oi* to a random site
Else If (*Pp* < Q)
Move the empty ant to a random site
End If
Else If ((agent carrying item *Oi*) and (site *Sj* empty))then
Compute the similarity of *f(Oi)* in this place
Calculate the *Pd* and generate a random number Q
If (*Pd* > Q) /* put-down rule*/
Drop item
Move the free ant to a randomly selected site
Else If (*Pd* < Q)
Move the ant with the item *Oi* to another random site
End If
End If
End For
End For

Fig. 2: The Lf Algorithm

It's very essential before moving to the next section in this paper to highlight the importance of ex-determining the various parameters and settings of LF. Let's set for example, the similarity coefficient, the picking up/dropping probability constants *Kp & Kd*, and the dimensions of the square panel nets….etc, will affect to a great extent the consequences of clustering process. As well as this, the noisy data within data collections will make the ant, while picking up these objects, unable to simply decide dropping them in their proper places which will decrease the converging speed of algorithm. Moreover, the arbitrary movement of ants which they make while searching to put or hold objects in the proper cells, will cause a waste of time and then decrease in the duration of performance. All these factors set an important question for reducing the distance between ants and objects, and then reduce the time of looking for objects, and decreasing the repeating times as well, that latter will lead to a better performance for the algorithm.

3 THE ALGORITHM EACOC

In this section we will explain the essential elements of the proposed model, which is distinct from the LF; In terms of effectiveness of both, fast convergences of the algorithm and the quality of clustering. We can divide these procedures: the digital environment, the behavior of ants and the degree of similarity

1-the digital environment:

We have found through the previous works that the objects, to be clustering, are being scattered arbitrarily on a square net of cells, as the digital ants move the objects from one cell to another in accordance with both holding and putting conducts. However, throughout the digital environment of this model, and to increase the converging speed of the algorithm, we turned each digital ant into a senseless one about the location of the others, or when they are free or even loaded, as well as the information related to the objects available out of the allowed visual space of the digital net associated with number of objects which should be collected depending on the next equation:

$$S_x = S_y = \gamma\sqrt{n} \qquad (8)$$

2-the conduct of the digital ants:

During a collecting process, ants conduct two main behaviors: picking up and dropping. In (LF) those two conducts were known as the picking up/dropping probability *Pp,Pd*. However, we have redefined these conducts in our model and reduced the probability as much as possible to increase the quality of collecting process and decreasing the execution duration. We've also used the exponent probability conversion function to increase the algorithm converging speed in comparison with the quadratic probability conversion function that we have reformed the picking up conduct to be given as following:

$$p_p = \left(\frac{1-e^{-cf}}{1+e^{-cf}}\right)^2 \qquad (9)$$

Where (C) is considered as constant affects the speed of algorithm converging, as well as the similarity degree between objects in the digital ants' colony. On the other hand, we have cancelled the probability concerning the dropping conduct by making digital ants to memorize the degree of similarity through which they carried the object. Thus if it is less than the similarity degree when putting it, they will move to an empty area arbitrarily to see if it's possible to put the carried object or not. In the other case, the ant will put the carried object in this cell to continue carrying other objects (it will clear his memory before doing that), as the degree of similarity function is given as following:

$$f(o_i) = \frac{1}{\pi\gamma^2}\sum(2 - d(o_i, o_j)) \qquad (10)$$

On the grounds that *d*(*Oi*,*Oj*) is not calculated according to the traditional Euclidean, but rather we have continued using Hyperbolic cosine function which has a high converging speed compared with Euclidean distance function, it is then given as:

$$d(o_i, o_j) = \cosh(\sum(o_i, o_j)^2) \qquad (11)$$

4 EXPERIMENTAL OUTCOMES AND PERFORMANCE ANALYSIS

In order to do the test of the performed work, efficiency and the quality of clustering results of the suggested model, we have used the standard data collection IRIS available in UIC [16], within the experimental collecting group. The dataset consists in 50 samples from each of three species of *Iris* flowers (*Iris setosa*, *Iris virginica* and *Iris versicolor*). Four features were measured from each sample; they are the length and the width of sepal and petal, in centimeters. Based on the combination of the four features, we have drawn a comparison between our experimental consequences we got through our simulation suggested algorithm with its equal at LF, which was used in [17], as we have implemented our model in VC++ 8.0. Whereas, concerning the parameters used in LF, they are as follows: similarity α=0.35, picking-up probability constant *kp* = 0.10, dropping probability constant *kd* = 0.15, the radius of the ant's visibility region γ = 3, the number of the ants = 6, the size of the two-dimensional space is 50×50, the number of the iterations = 1000000. But In EACOC, we have used the following settings: the number of the ants = 20, picking-up probability constant *kp* = 0.4, the radius of the ant's visibility region γ = 3, the number of the iterations = 1000. Figures (3),(4) show the initial distribution of data on the net at the starting point on both algorithm LF and EACOC. Figures (4) and (5) also show data distribution at the final point of collecting process in both cases, since (♦) represents (*Iris setosa*) data, (■) represents (*Iris virginica*) data, and (▲) represents (*Iris versicolor*) data.

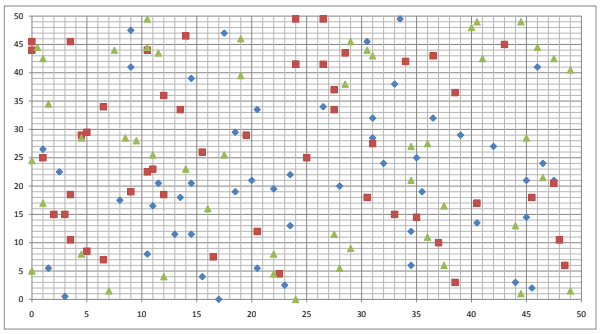

Fig.3: The distribution of data at the starting point (LF)

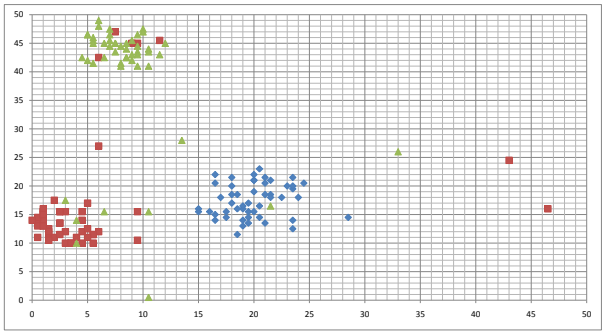

Fig.4: The distribution of data at the ending point (LF)

By looking at figure (4) we can notice that within the same cluster there are samples of other clusters, we can see also the existing of scattered samples without

being clustered in any group. However, by comparing (4) to (6) we can see the differentiation in EACOC which is widely better than that in LF, taking into consideration that the dimension in our model is less than that of LF model, in addition to that the number of digital working ants is more than that of LF, and this emphasizes the interactive relation between the ants in our digital environment, and leads to a more complex relation but at the same time the accuracy of clustering is much more better.

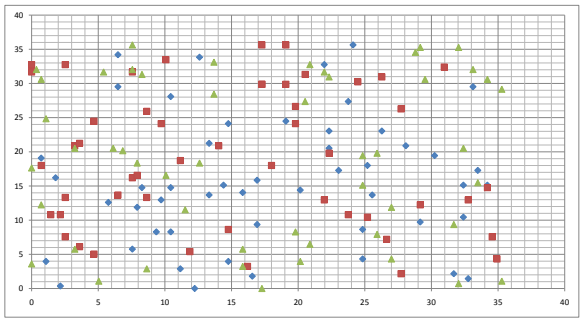

Fig.5: The distribution of data at the starting point (EACOC)

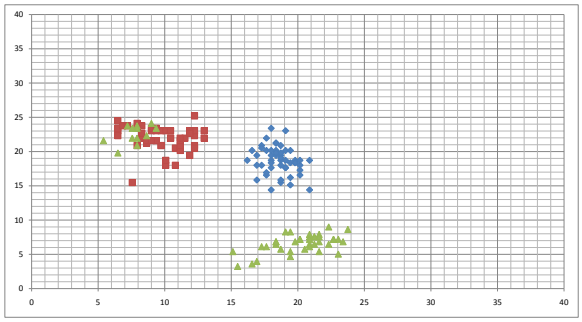

Fig.6: The distribution of data at the ending point (EACOC)

The following table shows as well the results of about (100) processes applied by both algorithm using the data set IRIS. By comparing these results we can realize the huge difference between our suggested model and LF depending on the number of repeating times, as well as the time needed to complete the clustering process just to get close results in both models.

	LF	EACOC
number of parameters ex-determining	5	3
number of repeating times	1000000	1500
number of cluster	6	3
average of errors	15	8
average of movements	4867	359
average of running time	48,96	1,12

5 CONCLUSION

The "efficient ant colony optimization clustering algorithm" **(**EACOC**)** is an extension of the classical ant algorithm LF, where it can deal with a large group of data more quickly while keeping the accuracy and efficiency at the same time. Thus, it overcame many obstacles in the LF as the large of iterative times, so that, the length of time of execution, which leads to the speed of convergence lower. Experimental results on clustering benchmarks demonstrate that EACOC algorithm has better performance and higher clustering quality. For future work, we can work on developing the performance of this kind of algorithm when dealing with other types of data as the case in the field of image processing.

The software developed has been implemented in the PTM3D system and applied to 3D segmentations and volume measurements from CT and MR images. The results have been used to plan and realize laparoscopic interventions [18].

This research work has been realized in the Radiological Imaging team of the LIMSI-CNRS laboratory at Orsay (France).

REFERENCES

[1] Kennedy J., Eberhart R. C., *Swarm Intelligence*, San Francisco: Morgan Kaufmann, 2001.

[2] Bonabeau E., M. Dorigo, G. Théraulaz. *Swarm Intelligence: From Natural to Artificial Systems*. Santa Fe Institute in the Sciences of the Complexity, Oxford University Press, New York, Oxford, 1999.

[3] Han, J. and Kamber, M., Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, 2006

[4] Jain, A.K., Murty, M.N., Flynn, P.J., "Data clustering: a review," ACM Computing Surveys, 31(3), pp. 264–323, 1999.

[5] Xu, R., Wunsch II, D., "Survey of Clustering Algorithms," IEEE Transactions on Neural Networks, 16(3), 645–678, 2005.

[6] Rokach, L., Maimon, O., Clustering Methods, In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, Springer,New York, 2005, pp. 321–352.

[7] Dorigo M, Maniezzo V, Colorni A, "The ant system: optimization by a colony of cooperating agents," IEEE Trans on SMC, 1996, 26(1): 28-41.

[8] Deneubourg, J. L., Goss, S., Franks N., "The dynamics of collective sorting robot-like ants and ant-like robots," Proc. of the 1st Conf. on Sim. of Adaptive Behavior, 1991, pp. 356-363.

[9] Lumer, E., Faieta, B., "Diversity and adaptation

in populations of clustering ants," In: Proc. of the 3rd International Conf. on Simulation of Adaptive Behavior: From Animals to Animals, 1994, pp. 501-508.

[10] Kuntz, P., Layzell, P., Snyder, D., "A colony of ant-like agents for partitioning in VLSI technology," Proceedings of the Fourth European Conference on Artificial Life, 1997, pp. 412–424.

[11] Kuntz, P., Snyder, D., "New results on ant-based heuristic for highlighting the organization of large graphs," Proceedings of the 1999 Congress or Evolutionary Computation, 1999, pp. 1451–1458.

[12] Kuntz, P., Snyers, D., Layzell, P., A Stochastic Heuristic for Visualizing Graph Clusters in a Bi-Dimensional Space Prior to Partitioning, Journal of Heuristics, 1999, pp. 327–351.

[13] Handl, J., Meyer, B., Improved Ant-Based Clustering and Sorting in a Document Retrieval Interface, PPSN VII, LNCS, 2002.

[14] Chrétien, L., "Organisation Spatiale du Matériel Provenant de l'excavation du nid chez Messor Barbarus et des Cadavres d'ouvrières chez Lasius niger (Hymenopterae: Formicidae)", Ph.D. dissertation, Université Libre de Bruxelles, 1996.

[15] Mahalanobis, P. C. (1936). On the generalised distance in statistics. Proc. Natl. Inst. Sci. India **12**, 49-55.

[16] UCI machine learning data-base http://archive.ics.uci.edu/ml/.

[17] L. Chen, X.H. Xu, Y.X. Chen, "An adaptive ant colony clustering algorithm," IEEE Machine Learning and Cybernetics, 2004, vol.3: 1387 – 1392.

[18] Angel Osorio, Juan-Antonio Galan, Julien Nauroy, Patricia Donars, Juan-Jose Lobato, Ines Navarro. "Real time planning, guidance and validation of surgical acts using 3D segmentations, augmented reality projections and surgical tools video tracking." *SPIE medical imaging. San Diego, USA, February 2010 (Com. N. 7625-80).*