Extended Firefly Algorithm and its Application in Unsupervised Classification of Mixed and Incomplete Data

Yenny Villuendas-Rey,[*1], José Felix Cabrera Venegas[1], and Jarvin Alberto Anton Vargas[1]

[1]University of Ciego de Ávila, Road to Morón km 9 ½, Ciego de Ávila, Cuba
{yenny, josefelix, janton@cav.uci.cu}

Abstract. Bioinspired Algorithms are ones of the most powerful metaheuristics used to solve optimization problems because its low computational cost and ability to explore a wide space of solutions in a little time. Unsupervised Classification may be comprehended as an optimization problem where the goal is to find the right data clustering among the numerous ways to do data clustering. In this paper we propose a new extension to Firefly Algorithm FA. The propose method is based in modifications to the original metaheuristic and the redefinition of artificial fireflies and the objective function for adjust the algorithm to solve a general problem. This approach was applied to Unsupervised Clasification with mixed and incomplete data. Experimental analysis with other algorithms using repository databases shows that our approach is able to find compact and separate clusters as well as to estimate the natural structuration of data.

Keywords. unsupervised classification, bioinspired metaheuristics, firefly algorithm (FA).

1 Introduction

Unsupervised Classification consists on obtaining a partition of the dataset in such way that the objects belonging to the same group be more similar to each other that with regard to the objects of other groups. In a general way, the algorithms of Unsupervised Classification are based on some approach that reflects how good it is a certain partition of the data [1].

Several authors have considered Unsupervised Classification as a problem of optimization, where the goal is to obtain the partition that maximizes or minimized the objective desired (the quality of the obtained structuring). Therefore, it has been carried out some proposals for the application of algorithms of optimization to the solution of this problem in particular. Contrary to most of the problems of optimization, where it is necessary to find an n-dimensional vector that satisfies the objective to optimize, in the case of Unsupervised Classification we have the space of solutions formed by all the possible partitions to obtain. In a similar way, it happens in other

Advances in Computer Science
Research in computing Science 81, 2014 pp. 47-57

domains of the Artificial Intelligence, such as the prototypes selection, rules generation, etc. [2]. Therefore, it is necessary to outline the problem of optimization in a general way, without assuming vectorial spaces.

Recently, bioinspired algorithms based in swarm intelligence have been applied successfully to the problem of the Unsupervised Classification [3], [4], [5], although only for numeric data, not being proposals for Mixed and Incomplete Data (MID). In this work, we will propose an extension to the Firefly Algorithm (FA) for solve general optimization problems, and we will apply it to the Unsupervised Classification of MID.

2 Extended Firefly Algorithm to obtain Groups in Mixed and Incomplete Data

Bioinspired Metaheuristics have proved their efficiency in Unsupervised Classification [6], [7], [8]. However, most algorithms only work only with numerical data. In this section we explain the metaheuristic Firefly Algorithm (FA) proposed by Xin-She Yang [9] for numerical optimization, and its extension to solve the problem of Unsupervised Classification where objects with mixed and incomplete data descriptions have to be arranged in groups.

2.1 New Metaheuristic based in an Extension to Firefly Algorithm

The social behavior of the fireflies has focused the attention of many computer scientists, basically regarding to the light they flash. The flashing light can be formulated in such a way that it is associated with the objective function to be optimized, which makes it possible to formulate new optimization algorithms. One of these techniques is a Firefly Algorithm (FA) for multimodal optimization applications developed by Xin-She Yang in 2009 [9]. FA was proposed originally for numerical optimization, it can be modified for obtaining of groups of objects in Unsupervised Classification of MID. These modifications are addressed below.

It is known that the selection of appropriate values for the parameters of the algorithms is crucial for a good performance of them. In this case, the algorithm FA consists of four fundamental parameters: β, γ (attractiveness and variation of the attractiveness) and the quantity of fireflies η. Investigations carried out by Lukasik and Zak [10] on the FA allows concluding that the best values for these parameters are: β =1, γ =1 and η varying between 15 and 50.

These conclusions allow simplifying the movement of the firefly defined in [10], as follows:

$$X_i = x_i + \beta(x_j - x_i + \alpha\varepsilon_i), \text{ with } \beta = 1$$

$$X_i = x_i + x_j - x_i + \alpha\varepsilon_i$$

$$X_i = x_j + \alpha\varepsilon_i \quad (1)$$

Therefore, the movement of firefly i transforms into a modification of the position of firefly j. In this idea we base ourselves to develop the extension of the algorithm. In a general way, a modification of the position of the brighter firefly can be considered like a "perturbation" of this firefly. With the objective of extending the original algorithm, we consider that each firefly consists, more than in a certain position, in a solution candidate to the problem of optimization we want to solve.

Furthermore, the characteristics of the fireflies in the original algorithm [10] are redefined in the following way:

1. Each firefly represents a candidate valid solution to the problem of optimization we want to solve.
2. All the fireflies are unisex, that means that a firefly can attract other fireflies without considerate its sex.
3. The attractiveness of the firefly is proportional to its brightness; in this case the less bright ones will be the result of a perturbation of brighter firefly, being the last one the more attractive. If there is not a firefly brighter that the current firefly, then the current firefly is perturbed because it is not attracted by any firefly.
4. The brightness of a firefly is affected or it is determined by the objective function that will be optimized.

Also, we assume that exists a way of "perturbing" this solution. Then, the "movement" of fireflies will be given for (2), instead of equation (1):

$$X_i = Perturb\ (x_j) \tag{2}$$

Each metaheuristic must implement a balance between search in the solutions space (exploration) and intensification of the good solutions (exploitation). In the particular case of the algorithms based on swarms of fireflies, the intensification of the good solutions is given by the realization of perturbations in the best fireflies. This procedure reduce the exploration of the search space by the perturbation of the best firefly we have found, reducing the possibilities to explore areas of the search space that are far from the area represented by the best fireflies.

To solve this problem, we include a new characteristic to the fireflies: their Time of Life. The Time of Life of firefly is a parameter of the algorithm, and it defines the number of iterations each firefly will "live". The age of the fireflies is initialized in 1 when the fireflies are generated initially, and it is increased each iteration. Every time that a firefly is perturbed, her age is again 1. Then, if a firefly is not perturbed during a number of iterations defined (it exceeds its time of life), we consider that this firefly dies, and it is replaced for other firefly generated according to the generation procedure used by the algorithm

This new characteristic is able to take out the algorithm of local optimal solutions and, at the same time, it improves the possibilities to explore big areas to the search space. Although, the modifications introduced in the metaheuristic allow a bigger exploration of the search space, it is possible to lose good solutions, even optimal solutions, due to the "death" of the fireflies. In Fig. 1, we show the pseudo code of the Firefly Algorithm with the modifications discussed in this section.

Extended Firefly Algorithm

Input: η: quantity of fireflies
A: fireflies attractiveness (objective function)
I: quantity of iterations
t: time of life of fireflies (age)

Output: *Best_F*: Best firefly (solution to the optimization problem)

Stage 1. Fireflies Inicialization (creating candidate solutions)

for $i = 1$ **to** η
  F_i = *Generate_Firefly()*
  $F_i.age = 1$
end for *i*
Hall_Fame = $arg\ \max_{i=1\ldots\eta}\{A(F_i)\}$

Stage 2. Shine of Fireflies (exploration the search space)

Iteration =1;
while (*Iteration* < *I*)
  for $i = 1$ **to** η
    for $j = 1$ **to** η
      if ($A(F_j) > A(F_i)$)
        F_i = *Perturb*(F_j)
        $F_i.age = 1$ **end if**
    end for *j*
    if ($F_i.age > t$)
      Consider_to_Fame(F_i)
      Replace_Firefly(F_i)
    else
      $F_i.age$++
    end if-else
  end for *i*
  Best_F = $arg\ max_{i=1\ldots\eta}\{A(F_i)\}$
  Best_F = *Perturb(Best_F)*
  Iteration++
end while
if (*A(Hall_Fame)* > *A*(*Best_F*))
  Best_F = *Hall_Fame* **end if**

Fig. 1. Pseudo code of extended firefly algorithm (EFA).

Another important feature is the fireflies die. For example, if we have a good solution in the 2nd iteration of the algorithm, and the time of life of the firefly is 3 iterations, this optimal solution won't be improved, and possibly, we will get lost the solution due to the firefly die. To solve this problem, we consider introducing a Hall of Fame in the metaheuristic. The Hall of Fame will be dedicated to store the information of the best firefly (solution of the problem) founded.

When a firefly dies, it will be considered to integrate the Hall of Fame (*Consider_to_Fame*). This means that the firefly in question (candidate) will be compared with the existent firefly in the Hall, and if it is better than the last one, the existent firefly in the Hall of the Fame will be replaced by the candidate firefly.

The new metaheuristc propose above, allow finding the solution of problems in a general way, not only restricted to numeric search spaces. Also, the introduction of Time of life and the Hall of Fame can improve the quality of the obtained solution. In the following section is analyzed the application of this metaheuristic in Unsupervised Classification of Mixed and Incomplete Data (MID).

2.2 Extended Firefly Algorithm applied to Unsupervised Classification of MID (EFAC)

The process of Unsupervised Classification can be seen as a combinatorial problem of optimization. In this case, the space of solutions is given by all the possible ways to create groups of objects, and the function to optimize is the quality of the obtained groups. We will call to the approach we will explain next: Extended Firefly Algorithm for Clustering (EFAC). With the goal of applying the proposed metaheuristic in Unsupervised Classification of MID, it is necessary firstly to define how the fireflies will be modeled and which attractiveness will be used (objective function). Also, we need to define how the firefly perturbation will be made.

First, we have to delimit the problem to the domain of Restrict Algorithms of Unsupervised Classification. This means that we need to know the value of k, the number of groups we want to obtain.

For solving our problem, a firefly will be a way to group the data objects (a candidate clustering). To model this clustering, each firefly represents the k centers of the k groups and is built in the following way:

$$F_i = (\bar{c}_1, \bar{c}_j \ \ldots, \bar{c}_k) \tag{3}$$

where represents the center of the group *j*, in the firefly *i* (*i* candidate clustering).

To allow the handling of MID, instead of using the average of the objects of the group like center, we select as center of a group the object that minimizes the disimilarity with rest of objects of its group:

$$\bar{c}_j = arg\, min_{x,y \epsilon C_j}\{d(x,y)\} \tag{4}$$

To compute the distance between two objects $d(x,y)$ with MID attributes, we used the HEOM dissimilarity (equation 5) proposed by Wilson and Martínez [11].

$$HEOM(x,y) = \sqrt{\sum_{a=1}^{m} d_a(x_a, y_a)}$$

$$d_a = \begin{cases} 1 & unknown\ attribute \\ overlap(x_a, y_a) & nominal\ attribute \\ diff(x_a, y_a) & otherwise \end{cases}$$

$$overlap(x_a, y_a) = \begin{cases} 0 & if\ x_a = y_a \\ 1 & elsewhere \end{cases}$$

$$diff(x_a, y_a) = |x_a - y_a| / (max_a - min_a) \quad (5)$$

max_a and min_a are the maximum and minimum values of attribute a, respectively.

Other important element in the algorithm is the attractiveness or objective function. This function measures the fitness of each clustering (solution) and permit to know which clustering is the best. The objective function we use in our investigation is the Dunn index (equation 6), defined in [12]. The Dunn index for a clustering is the quotient between the smallest distance among two groups, and the size of the biggest group. A high value of Dunn index for a clustering means we have more compact and more separate groups [13].

$$D = \frac{\min\limits_{i=1..k,\, j=1..k,\, i \neq j} \{d(c_i, c_j)\}}{\max\limits_{i=1..k} \{\Delta(c_i)\}} \quad (6)$$

Another element of the algorithm is the generation of fireflies. In our case, we use a random generation process. We select of the database of objects *k* elements that will be the centers of the groups. This strategy of random selection of the centers of the groups allows exploring a wide area of the search space, and it allows maintaining the diversity in the group of fireflies.

Finally, we defined how to perturb the fireflies. For this, we develop a method to produce a new solutions form a current solution inspires by the strategy of mutation of the Genetic Algorithm proposed for [7]. Our perturbation consists on replacing randomly one of the centers of the groups for another object in a random way.

The selection of the center that will be replaced is realized randomly. For this, a random number is generated indicating what center will be changed. Then, another random number is generated representing the object of the dataset that will replace the selected center. This strategy allows the exploration of wide areas of the search space, and it avoids a premature convergence of the algorithm.

The Fig. 2 shows an example of a firefly perturbation. In the figure, we represent the original firefly and below, the perturbed firefly. In de perturbed, the center of the second group is replaced randomly by another object of the group.

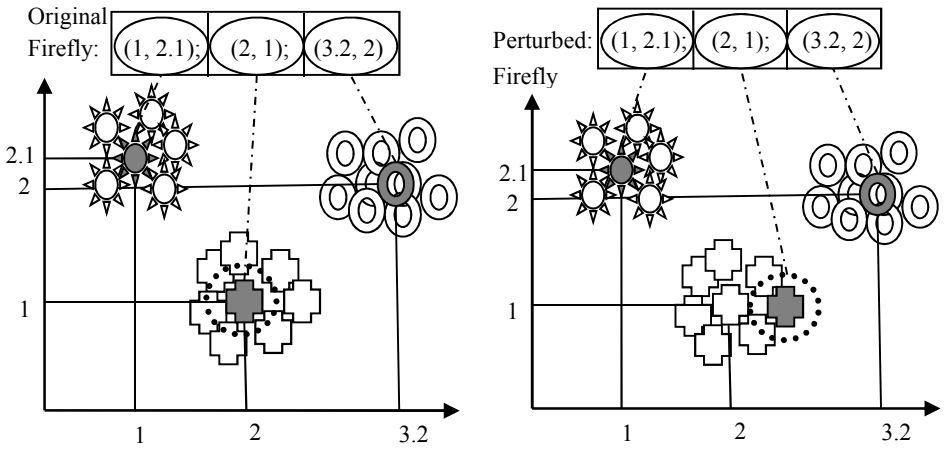

Fig. 2. Representation of a firefly perturbation.

3 Experimentation and Results

To check the performance of the proposed algorithm (EFAC) we developed experiments with six algorithms used in Unsupervised Classification of MID. These algorithms were divided in two groups. In the first we have the algorithms based in kmeans: KP [14], KMSF [15], AD2011 [16]. In de second group we have algorithms that use hierarchical and metaheuristic approaches: HIMIC [17], CEBMDC [18] and AGKA [7]. In the comparison were used seven data sets of Mixed and Incomplete Data (MID) from the UCI Repository [19], in Table 1 are shown the description of each data set.

Table 1. Data sets description.

Data sets	Categorical Features	Numerical Features	Classes
colic	15	7	2
dermatology	1	33	6
heart-c	7	6	5
hepatitis	13	6	2
labor	6	8	2
lymph	15	3	4
tae	2	3	3

To measure the obtained results we used the classes of the data sets as the real clustering. For each obtained clustering we calculated its Entropy [20]. Entropy is an external validity index defined by:

$$E = -\sum_i \frac{|C_i|}{|O|} \sum_j \frac{\left|\left\{o \in C_i \mid \alpha(o) = l_j\right\}\right|}{|C_i|} \log\left(\frac{\left|\left\{o \in C_i \mid \alpha(o) = l_j\right\}\right|}{|C_i|}\right) \quad (9)$$

Where O is the set of objects (data set), C is the obtained clustering, C_i is the group i, l_j is the class j and α (o) it is the class of the object o.

Entropy gives a measure of the difference between the groups obtained by an unsupervised classifier and the original classes in a data set. According to the above-mentioned Entropy measures the grade of disorganization of the original groups (classes in the data set) with regard to obtained groups (clustering results). Therefore, when the value of Entropy is smaller, the groups will be more similar to the data set classes.

We applied the different algorithms for each data set. Then we computed the Entropy index of each obtained clustering. In each algorithm, the quantity of groups was established as the quantity of classes for each data set. The dissimilarity used in the experiments for all the algorithms was HEOM [11] (equation 8). This dissimilarity has been applied in several experimental studies about MID [21].

In Table 2 we show the results according to Entropy, the best results are highlighted in bold.

Table 2. Entropy Results of EFAC vs. the rest of algorithms.

Data sets	KP	KMSF	AD2011	HIMIC	CBMDC	AGKA	EFAC
colic	0.9658	0.9344	0.9503	0.9475	0.9488	0.9451	**0.8978**
dermatology	2.3625	1.731	2.4326	2.4326	2.306	2.4256	**0.6976**
heart-c	0.995	0.996	0.9943	0.9943	0.991	0.9939	**0.9687**
hepatitis	0.7203	**0.5643**	0.7346	0.7346	0.7381	0.7512	0.7344
labor	0.9407	0.7601	0.9348	0.9348	0.9077	0.9456	**0.6486**
lymph	1.7508	1.7856	1.8813	1.703	1.6721	1.7999	**0.9124**
tae	1.5834	1.5818	1.5845	1.5821	**1.5459**	1.5815	1.5515

For clarify these results about the performance of the algorithms, it was necessary find out if they have or haven't significant differences in their performing. For it, we use the methodology recommended by Demsar for the comparison of classifiers in multiple databases [22].

Firstly, we established α =0.05, for a 95% of confidence. Then, for each pair (EFAC vs. Algorithm), we fix the following hypotheses: H0: In the performance of the algorithm EFAC and the other algorithm don't exist significant differences, and H1: In the performance of the algorithm EFAC and that of the other algorithm exist significant differences.

After that, we apply a Wilcoxon test to the results obtained by each pair of algorithms (the algorithm EFAC with each one of the other algorithms). The results of the test are presented in Table 3. Each column show the probability of the Wilcoxon test, and the times the EFAC won, lose or tie with respect other. In each case, if the obtained probability is smaller than 0.05, it is possible to reject the null hypothesis.

Table 3. Wilcoxon test on Entropy.

EFAC vs	Win – Loss – Tie	Probability
KP	7-0-0	**0.018**
KMSF	6-1-0	0.128
AD2011	7-0-0	**0.018**
HIMIC	7-0-0	**0.043**
CBMDC	6-1-0	**0.018**
AGKA	7-0-0	**0.018**

Considering these results we may conclude that the proposed algorithm (EFAC) exceeds in performance to the rest of algorithms except the method KMSF. In this case, the experiment was not enough to say if exist or don't differences between the algorithms.

These results allow saying that metaheuristics based in FA have promissory results for unsupervised classification of mixed and incomplete data. However, it is necessary accomplishing more extensive future experiments.

4 Conclusions

In this paper a new extension to Firefly Algorithm FA is proposed. Our extension, denominated Extended Firefly Algorithm, enlarges the concept of artificial firefly, allowing the modeling of other optimization problems defined in non-numeric domains. The algorithm was applied for the unsupervised classification of mixed and incomplete data. Experimental results allow affirming that the proposed algorithm achieves similar or superior performance respect other methods reported in specialized literature. Therefore we may conclude that the proposed algorithm is able to find the natural structuration of data.

References

1. Handl, J., Knowles, J.: An evolutionary approach for multiobjective clustering. IEEE Transactions on Evolutionary Computation, Vol.11, No.1, pp. 56 – 76 (2007).
2. Ahn, H., Kim, K. J.: Global optimization of case-based reasoning for breast cytology diagnosis. Expert Systems with Applications, Vol.36, pp. 724 - 734 (2009).
3. Prior, A. K. F., Nunes de Castro, L.: The proposal of two bio-inspired algorithms for text clustering. Learning and Nonlinear Models - Revista da Sociedade Brasileira de Redes Neurais (SBRN), Vol.6, No.1, pp. 29 – 43 (2008).
4. Charalambous, C., Cui, S.: A Bio-Inspired distributed clustering algorithm for wireless sensor networks. WIXON'08, ACM (2008).
5. Ahmad, A., Dey, L.: A k-means clustering algorithm for mixed numerical and categorical data. Data & Knowledge Engeneering, Vol.63, pp. 503 – 527 (2007).

6. Hassanzadeh, T., Meybodi, M. R.: A new hybrid approach for data clustering using firefly algorithm and K-means, in Artificial Intelligence and Signal Processing (AISP). 16th CSI International Symposium on. IEEE (2012).
7. Roy, D. K., Sharma, L. K.: Genetic k-means clustering algorithm for mixed numeric and categorical datasets. International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.2 (2010).
8. Errecalde, M. L., Ingaramo, D. A.: A new AntTree-based algorithm for clustering short-text corpora. JCS&T, Vol.10, No.1 (2010).
9. Yang, X. S. Firefly algorithms for multimodal optimization. Lecture Notes in Computer Sciences, pp. 169 – 178 (2009).
10. Lukasik, S., Zak, S.: Firefly Algorithm for Continuous Constrained Optimization Tasks (2009).
11. Wilson, R. D., Martinez, T.R.: Improved Heterogeneous Distance Functions. Journal of Artificial Intelligence Research, Vol.6, pp. 1 – 34 (1997).
12. Azuaje, F.: A cluster validity framework for genome expression data. Bioinformatics, Vol.18, pp. 319 – 320 (2002).
13. Brun, M.: Model-based evaluation of clustering validation measures. Pattern Recognition, Vol.40, pp. 807 – 824 (2007).
14. Huang, Z.: Clustering large data sets with numeric and categorical values. 1rst Pacific - Asia Conference on Knowledge discovery and Data Mining (1997).
15. García-Serrano, J.R., Martínez-Trinidad, J.F.: Extension to c-means algorithm for the use of similarity functions. 3rd European Conference on Principles of Data Mining and Knowledge Discovery, Prague (1999).
16. Ahmad, A., Dey, L.: A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical data. Pattern Recognition Letters, Vol.32, pp. 1062 – 1069 (2011).
17. Ahmed, R.A.: HIMIC: A Hierarchical Mixed Type Data Clustering Algorithm. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.6369&rep=rep1&type=pdf (2005).
18. He, Z., Xu, X., Deng, S.: Clustering mixed numeric and categorical data: A cluster ensemble approach. CoRR, http://arix.org/abs/cs/0509011 (2005).
19. Merz, C. J., Murphy, P.M.: UCI Repository of Machine Learning Databases, University of California in Irvine, Department of Information and Computer Science, Irvine (1998).
20. Hsu, C., Chen, C., Su, Y.: Hierarchical clustering of mixed data based on distance hierarchy. Information Sciences, Vol.177: pp. 4474 – 4492 (2007).
21. Villuendas-Rey, Y.: Selecting features and objects for mixed and incomplete data. 13th Iberoamerican Congress in Pattern Recognition. CIARP 2006, LNCS 5197, Springer Heidelberg, La Habana, pp. 381 – 388 (2008).
22. Demsar, J.: Statistical comparison of classifiers over multiple datasets. The Journal of Machine Learning Research, Vol.7, pp. 1 -30 (2006).