A CLUSTERING MODEL USING ARTIFICIAL ANTS

Hanene Azzag, Gilles Venturini

Laboratoire d'Informatique, Ecole Polytechnique de l'Université de Tours,
Ecole Polytechnique de l'Université de Tours - Département Informatique
64, Avenue Jean Portalis, 37200 Tours, France.
Tél : +33 2 47 36 14 14, Fax: +33 2 47 36 14 22
hanene.azzag@etu.univ-tours.fr, venturini@univ-tours.fr

Abstract:

In this paper we will present a new clustering algorithm for unsupervised learning. It is inspired from the self-assembling behavior observed in real ants where ants progressively become attached to an existing support and then successively to other attached ants. The artificial ants that we have defined will similarly build a tree. Each ant represents one data. The way ants move and build this tree depends on the similarity between the data. We have compared our results to those obtained by the k-means algorithm and by AntClass on numerical databases (either artificial or real,). We show that AntTree significantly improves the clustering process.

Introduction:

Many data mining systems often require the use of a clustering algorithm. Natural systems have evolved in order to solve many problems that can be related to clustering. Different species have developed social behaviors to tackle the problem of gathering objects or individuals. For instance, we can cite the brood sorting or cemetery organization of ants [5] or the collective movements in different species such as the ability of bacteria to form surprising spatial patterns and aggregations [4]. Many researchers in computer science have been inspired by real ants [8] and have defined artificial ants paradigms for dealing with optimization or machine learning problems [3].
In this paper, we propose the adaptation of a new biological model which, as far as we know, has never been used before to solve computer science problems.
We model the ability of ants to build live structures with their bodies [12] in order to discover, in a distributed and unsupervised way, a tree-structured organization of the data set. This hierarchical structure can be interpreted in several ways: it can be seen as a partitioning of the data (that we will compare with other clustering methods), or it can be used for data visualization purposes, as in hierarchical clustering [9].

A model for self-assembling behaviors in real ants:

Real ants provide a stimulating self-organization model for the clustering problem. Previous models involve the ability of ants to sort objects [13][10][6][14] or to build a colonial odor [11]. Here, we consider another biologically observed behavior: Ants are able to build mechanical structures by a self-assembling behavior. This can be for instance the formation of drops constituted of ants [15], or the building of chains by ants with their bodies in order to link leaves together [12]. These types of self-assembly behavior have been observed with ***Linepithema humiles*** Argentina ants and African ants of gender ***Oecophylla longinoda***. The goal of drop structures built ***L. humiles*** is today still obscure. This ability has been recently experimentally demonstrated [15]. The drop can sometimes fall down. For ***Oecophylla longinoda*** ants, it can be observed that two types of chains are built: for crossing an empty space, and on the other hand for building their nest [12]. In both cases, crossing chains and building chains, these structures disaggregate after a given time.
From these self-assembly behaviors, we can extract properties that will constitute the framework of our algorithm:

- Ants build this type of live structures starting from a fixed support (stem, leaf,…),
- Ants can move on this structure whilst it is being currently built,
- Ants can cling anywhere on the structure because every position can be reached. Nevertheless, in the formation of chains for example, ants will preferably fix themselves at the end of the chain, because they are attracted by gravity or by the object to reach,
- The majority of ants which constitute the structure can be blocked without any possibility of displacement. For example, in the case of a chain of ants,

this corresponds to ants placed in the middle of the chain,

- Some ants (a number generally much more reduced than for blocked ants considered in the previous point) are connected to the structure but with a link which they maintain by themselves, they can thus be detached from the structure whenever they want. In the case of a chain of ants, this corresponds to the ants placed at the end of the chain,
- We can observe a phenomenon of growth but also of decrease of the structure.

In general, the motivation for using bio-inspired clustering techniques is twofold: they can avoid local minima thanks to their probabilistic behavior and they can produce high quality results without any prior knowledge of data structure (such as the number of clusters or an initial partition). In this work, in addition to these motivations, we are especially interested in showing that this new biological model may be a promising technique for achieving parallel tree-based clustering.

The AntTree algorithm: General principles

To obtain a partitioning of the data, we build a tree where nodes represent data and where edges remain to be discovered. One should notice that this tree will not be strictly equivalent to a dendogram as used in standard hierarchical clustering techniques:
Each node in our tree will correspond to one data while this is not the case in general for dendrograms, where data only correspond to leaves [9].

We consider in the following that each data can be described by any representation language, provided that there exists a similarity measure between two data. In the following, we denote this similarity
measure by *Sim(i, j)* which gives, for a couple of data *(di, dj)*, *i* in *[1,N]*, j in *[1,N]*, a value in *[0,1]* (*N* indicates the number of data). *0* means that *di* and *dj* are totally different and *1* means that they are identical. We do not need any additional hypothesis about data representation.

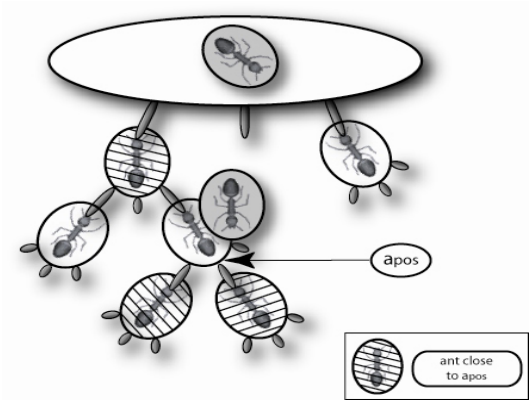

Figure 1

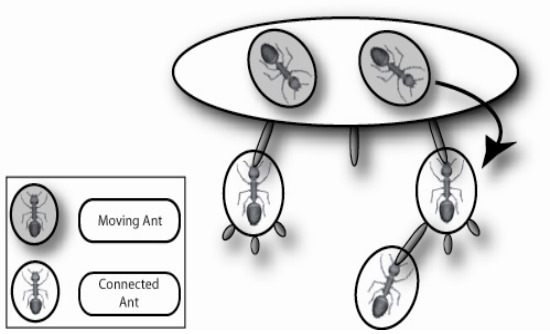

Figure 2

The main principles of our algorithm called AntTree are the following (see figure 1): each ant represents a node of the tree to be assembled, i.e. a data to be clustered. On the basis of a root materialized by a fictitious node a_0 which represents the support on which the tree will be built, ants will gradually fix themselves on this initial node, and then successively on the ants fixed to this node, and so on until all ants are attached to the structure.
All these moves and these fixings depend on the value returned by *Sim(i, j),* and on the local neighborhood of the moving ants. Thus, for each ant *ai*, *i* in *[1, N]* we need to define the following concepts:

- the outgoing link of *ai* is the link that *ai* can maintain toward another ant,
- the incoming links of *ai* are the links that the other ants maintain toward *ai*, these bonds can be the legs of the ant,
- the data *di* represented by *ai*,
- a similarity threshold *T_Sim_ai* and a dissimilarity threshold *T_Dissim_ai* that will be locally updated by *ai*.

During the assembly of the structure, each ant *ai* will be either:

- **Moving on the tree**: *ai* moving over the support a_0 or over another ant denoted by *a_pos*, but ai is not connected to the structure (see the ants colored in gray on figure 1} and 2). It is thus completely free to move on the support or toward another ant within its neighborhood. If *a_pos* denotes the ant where *ai* is located on, then *ai* will move randomly to any immediate neighbors of *a_pos* in the tree (considering in this case that edges are undirected, as shown in figure 2),

- **Connected to the tree**: *ai* can no be longer released anymore from the structure. Moreover we will consider the fact that each ant has only one outgoing link toward other ants and cannot have more than *L_max* incoming links from other ants. This enables us to build a tree having a maximum degree of *L_max* links for any given node (see figure 2).

Experimental results:

In order to evaluate and compare results obtained by AntTree, we have used databases where data is represented with numerical attributes (one data is a vector of real).
Two kinds of databases are used: artificially generated ones (Art1,..., Art8) and real ones (Iris, Wine, Glass, Pima, Soybean and Thyroid) from the Machine Learning Repository [1]. We have also used data from the CE.R.I.E.S. Research center on healthy human skin domain [7]. These databases are supervised ones because for each data we know the cluster it belongs to. In this way, it is possible to evaluate the clusters we obtain with respect to these real classes (but of course real classes are not given to the clustering methods).
We denote by *ki* the known class number for data *di* and by *k'i* the class number computed by a clustering method. *K* corresponds to the real number of classes and *K'* corresponds to the number of classes found by one of our methods. We have used the following classification error measure:

$$Ec = \frac{2}{N(N-1)} \sum_{(i,j)\in\{1,\ldots,N\}^2, i<j} \epsilon_{ij} \qquad (1)$$

where :

$$\epsilon_{ij} = \begin{cases} 0 & if\,(k_i = k_j \wedge k'_i = k'_j) \vee (k_i \neq k_j \wedge k'_i \neq k'_j) \\ 1 & else \end{cases} \qquad (2)$$

Database	AntTree	
	$Ec\,[\sigma_{Ec}]$	$K'\,[\sigma_{K'}]$
Art1	0.75 [0.00]	1 [0.00]
Art2	0.50 [0.00]	1 [0.00]
Art3	0.58 [0.00]	1 [0.00]
Art4	0.43 [0.00]	3 [0.00]
Art5	0.36 [0.00]	2 [0.00]
Art6	0.53 [0.00]	1 [0.00]
Art7	0.54 [0.00]	4 [0.00]
Art8	0.61 [0.00]	5 [0.00]
Iris	0.67 [0.00]	1 [0.00]
Wine	0.65 [0.00]	2 [0.00]
Glass	0.71 [0.00]	3 [0.00]
Pima	0.45 [0.00]	1 [0.00]
Soybean	0.15 [0.00]	3 [0.00]
Thyroid	0.38 [0.00]	3 [0.00]
CERIES	0.76 [0.00]	2 [0.00]

Table 1: Results obtained by AntTree. *Ec* corresponds to the averaged classification error on 50 runs and *K'* to the averaged number of classes, *σ_Ec* and *σ _K'* are the corresponding standard deviations, and *N* a number of data.

database	10-Means		AntClass	
	$Ec\,[\sigma_{Ec}]$	$K'\,[\sigma_{K'}]$	$Ec\,[\sigma_{Ec}]$	$K'\,[\sigma_{K'}]$
Art1	0.18 [0.01]	8.58 [0.98]	0.15 [0.05]	4.22 [1.15]
Art2	0.38 [0.01]	8.52 [0.96]	0.41 [0.01]	12.32 [2.01]
Art3	0.31 [0.01]	8.28 [0.96]	0.35 [0.01]	14.66 [2.68]
Art4	0.32 [0.02]	6.38 [0.75]	0.29 [0.23]	1.68 [0.84]
Art5	0.08 [0.01]	8.82 [0.91]	0.08 [0.01]	11.36 [1.94]
Art6	0.10 [0.02]	8.46 [1.08]	0.11 [0.13]	3.74 [1.38]
Art7	0.87 [0.02]	7.76 [1.03]	0.17 [0.24]	1.38 [0.60]
Art8	0.88 [0.01]	8.78 [0.83]	0.92 [0.01]	13.06 [2.18]
Iris	0.18 [0.03]	7.12 [1.11]	0.19 [0.08]	3.52 [1.39]
wine	0.27 [0.01]	9.64 [0.52]	0.51 [0.11]	6.46 [2.10]
Glass	0.29 [0.02]	9.44 [0.70]	0.40 [0.06]	5.60 [2.01]
Pima	0.50 [0.01]	9.90 [0.36]	0.47 [0.02]	6.10 [1.84]
Soybean	0.13 [0.02]	8.82 [0.97]	0.54 [0.17]	1.60 [0.49]
Thyroid	0.42 [0.02]	9.56 [0.57]	0.22 [0.09]	5.84 [1.33]
CERIES	0.11 [0.01]	9.38 [0.63]	0.27 [0.15]	3.40 [1.06]

Table 2: Results obtained with 10-means and AntClass algorithms.

We have compared AntTree with other clustering algorithms: AntClass [14], a clustering algorithm inspired by a colony of artificial ants, and the K-means algorithm initialized with 10 randomly generated initial partitions (the data used for experimentation do not contain more than 10 clusters).
Table 1 and 2 shows the results obtained for the 10-means, AntClass and AntTree. We can see that AntTree gives an averaged error which is lower than AntClass for Art2, Art3, Art4, Art8, Pima and soybean and almost similar for Art1, glass, thyroid. Moreover, for the majority of the databases, the number of clusters found by AntTree is closer to the number of real classes than the number found by AntClass (10 databases out of 15).
AntTree is also better than 10-means method for Art2, Art3, Art4, Art7, Art8, Pima, soybean and thyroid. Moreover, the number of classes found by AntTree is also better (14 databases out of 15) for these second results. According to the standard deviations, we can also note that AntTree is more precise than AntClass and 10-means.

Conclusion

We have described a new algorithm directly inspired from the ants self-assembly behavior and its application to the unsupervised learning problem. This method has been successfully compared with the k-means and the AntClass algorithms, those results are extremely encouraging and the main perspective of this work is to keep on studying this promising model.

References:

[1] BLAKE C., MERZ C., “UCI Repository of machine learning databases”, 1998.

[2] LIONI A., SAUWENS C., THERAULAZ G., DENEUBOURG J.-L., “The dynamics of chain formation in Oecophylla longinoda”, Journal of Insect Behavior,Conference, vol. 14, 2001, p. 679-696.

[3] BONABEAU E., DORIGO M., THERAULAZ G., Swarm Intelligence: From Natural to Artificial Systems, Oxford University Press, New York, 1999.

[4] CAMAZINE S., DENEUBOURG J.-L., FRANKS N. R., SNEYD J., THERAULAZ G., BONABEAU E., Self-Organization in Biological Systems, Princeton University Press, 2001.

[5] FRANKS N.-R., SENDOVA-FRANKS A., “Brood sorting by ants: distributing the workload over the work surface”, Behav. Ecol. Sociobiol, vol. 30, 1992, p. 109-123.

[6] GOSS S., DENEUBOURG J.-L., “Harvesting by a group of robots”, VARELA, Ed., Proceedings of the First European Conference on Artificial Life, Sydney, Australia, 1991, Toward a Practice of Autonomous Systems, p. 195–204.

[7] GUINOT C., MALVY D. J.-M., MORIZOT F., TENENHAUS M., LATREILLE J., LOPEZ S., TSCHACHLER E.,DUBERTRET L., “Classification of healthy human facial skin”, Textbook of Cosmetic Dermatology Third edition, 2003.

[8] HOLLDOBLER B., WILSON E.-O., The Ants, Springer Verlag, Berlin, 1990.

[9] JAIN A. K., MURTY M. N., FLYNN P. J., “Data clustering: a review”, ACM Computing Surveys, vol. 31, num. 3, 1999, p. 264–323.

[10] KUNTZ P., SNYERS D., LAYZELL P., “A stochastic heuristic for visualising graph clusters in a bi-dimensional space prior to partitioning”, Journal of Heuristics, 1999.

[11] LABROCHE N., MONMARCH N., VENTURINI G., “A new clustering algorithm based on the chemical recognition system of ants”, HARMELEN F., Ed., Proceedings of the 15th European Conference on Artificial Intelligence, IOS Press.

[12] LIONI A., SAUWENS C., THERAULAZ G., DENEUBOURG J.-L., “The dynamics of chain formation in Oecophylla longinoda”, Journal of Insect Behavior, vol. 14, 2001, p. 679-696.

[13] LUMER E., FAIETA B., “Diversity and Adaptation in Populations of Clustering Ants”, 3th Conference on simulation and adaptive behavior: from animals to animats, Cambridge, 1994, p. 501–508.

[14] Monmarché N., “On data clustering with artificials ants”, FREITAS A., Ed., GECCO-99 Workshop on Data Mining with Evolutionary Algorithms, Florida, 1999, p. 23-26.

[15] THERAULAZ G., BONABEAU E., SAUWENS C., DENEUBOURG J.-L., LIONI A., LIBERT F., PASSERA L., SOL R.-V., “Model of droplet formation and dynamics in the Argentine ant (Linepithema humile Mayr)”, Bulletin of Mathematical Biology, vol. 63, 2001.