Dealing with Noise in Ant-based Clustering

Daniela Zaharie
Department of Computer Science
West University of Timişoara
bv. V. Pârvan, no. 4, 300223 Timişoara, Romania
dzaharie@info.uvt.ro

Flavia Zamfirache
Department of Computer Science
West University of Timişoara
bv. V. Pârvan, no. 4, 300223 Timişoara, Romania
zflavia@info.uvt.ro

Abstract- Separating the noise from data in a clustering process is an important issue in practical applications. Various algorithms, most of them based on density functions approaches, have been developed lately. The aim of this work is to analyze the ability of an ant-based clustering algorithm (AntClust) to deal with noise. The basic idea of the approach is to extend the information carried by an ant with an information concerning the density of data in its neighborhood. Experiments on some synthetic test data suggest that this approach could ensure the separation of noise from data without significantly increasing the algorithm's complexity.

1 Introduction

In data analysis the clusters are seen as homogeneous groups of similar data and the main aim of a clustering task is to divide the data in groups such that the data in a group are sufficiently similar while the data belonging to different groups are sufficiently dissimilar. In practical applications the data can contain both useful items and irrelevant ones, considered as noise. If the data are noisy, the results of a clustering process can be altered, thus the noise should be identified and separated from the useful data.

In order to separate the noise from useful data a criterion which allows the discrimination between them should be used. Such a criterion is based on a measure of the data density. With respect to this criterion, clusters could be seen as dense regions of data while the noise corresponds to regions with low density of data. It is difficult to define an absolute threshold for the density values because the density corresponding to noise in a set of data could be similar to the density corresponding to clusters for another set of data. Rather a relative threshold depending, for instance, on some order statistics should be used.

Based on the idea of using a density measure to separate the noise from data and to identify the clusters, different approaches have been developed. Some of them are: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [Ester, 1996], DENCLUE (Density Clustering) [Hinneburg, 1998] and UNC (Unsupervised Niche Clustering) [Nasraoui, 2004].

In DBSCAN the measure of density is the number of points within a certain distance of each other. Based on this measure is defined the concept of *density-based connectivity* and the points are separated in *core points* (these represents the main portion of clusters), *border points* (these delimitate the clusters from the background) and other points (these represent the noise). This approach allows identifying clusters of arbitrary shapes and ignoring background noise.

DENCLUE uses a different measure of density constructed by using some *influence functions* which measure the influence that data have in their neighborhood. A common influence function is the gaussian one, i.e. the influence a data x has on a data y is expressed as $f(x, y) = \exp(-d(x, y)^2/(2\sigma^2))$ where $d(x, y)$ is a dissimilarity measure. The density in x can then be defined as $D(x) = \sum_{y\in\mathcal{N}_x} f(x, y)$, $\mathcal{N}_x$ denoting a neighborhood of x. The clusters are identified by the so-called density attractors defined as local maxima of the density function, $D(x)$. For a given data the corresponding attractor is determined by using gradient information. In order to identify arbitrarily shaped clusters the concept of high density path (similar with density based connectivity in DBSCAN) is used. The decision if the density is high or low is based on a threshold ξ. This threshold is used also to separate the data from noise and as long as its value is adequately chosen it leads to a successful cleaning of data. The ability of DENCLUE to identify the true clusters is highly dependent on σ and ξ.

A similar density function, based on gaussian influence functions, was used by Nasraoui at al. in designing the UNC algorithm. However they used a different approach in identifying the clusters. Their approach is based on using a genetic algorithm with a niching mechanism (deterministic crowding) to find the local maxima of the density functions. These local maxima represent unbiased estimates of the clusters centers. The clusters are supposed to be hyperellipsoidal and the algorithm produces the clusters centers and the parameters defining their size and orientation. The classification of data is then based on computing the distance between each data and each cluster center and to decide which cluster the data belongs to. The data not assigned to a cluster are considered to be noise.

The behavior of ant colonies inspired the development of various clustering techniques. Different approaches are based on different aspects of real ants behavior: (i) cemetery organization and larval sorting; (ii) chemical recognition of nestmates.

The first approach has been proposed by [Deneubourg, 1991] in order to solve tasks in robotics and adapted to data clustering by [Lumer, 1994]. A thorough analysis is presented in [Handl, 2003] where different improvements are also proposed. The basic idea of this approach is to place the data on a bi-dimensional grid such that the intrinsic clusters structure is reflected by the spatial arrangement of data. The process is carried on by a set of agents (ants) which wander on this grid and pick up and drop data according to some probabilities which are determined by the similarities between data.

The second approach has been proposed in [Labroche, 2002] and is inspired by the existence of an individual chemical odor. An ant uses this odor to recognize its nestmates. By putting into correspondence a nest to a cluster, a chemical odor to a cluster label and by modelling some behavioral rules, in [Labroche, 2002] is proposed an ant-based clustering algorithm called AntClust. Unlike the previous approach where an ant picks up and drops different data (and the number of ants is significantly smaller than the cardinality of the data set) in AntClust the ants are themselves the data (each ant is associated to a given data, so the number of ants is exactly the number of data to be processed). Thus, notions as ants and data or nests and clusters are interchangeable.

The clustering process simulates the process of generating new nests, accepting ants in nests and reorganizing nests. All of these are made by simulating meetings between ants when they confront their chemical labels and as a result decide that they belong to the same nest or not. The key elements of the clustering process are some individual adaptive parameters: (i) the acceptance threshold of an ant; (ii) the ant's perception of the size of its current nest; (iii) the ant's perception of the acceptance degree in its current nest. These values are used both in the iterative process of meetings during which they are adapted and in the final step of refining the clusters.

The aim of this paper is to analyze the ability of AntClust to deal with noisy data. With this respect we analyzed if the existing parameters give us enough information to separate the noise from data and also analyzed the effectiveness of introducing a new adaptive parameter related to the density of data. The rest of the paper is organized as follows. Section 2 gives a more detailed description of AntClust and presents results on identifying clusters in noisy data for some synthetic test data. In Section 3 we present a slight modification of AntClust which uses also information on density and analyze its behavior. Finally Section 4 concludes the work.

2 Description of AntClust

As stated before, AntClust, the algorithm proposed in [Labroche, 2002], simulates the so-called "colonial closure" phenomenon in ants colonies. This phenomenon is based on some chemical odors the ants possess and which allow them to recognize the difference between nestmates and intruders. Each ant has its own view on the colony odor. This is continuously updated. Starting from these ideas, Labroche et al. proposed a model of an artificial ant able to participate to the clustering of a set of data.

Let $\{x_1, \ldots, x_m\}$ be the set of data to be processed. Then a set of m ants are used. Each ant, i, has the following characteristics:

- An associated data, x_i. This is the unique element which is not modified during the clustering process.
- A label, L_i. This label is a natural number which identifies a cluster. Initially it is set to 0 (meaning that the data has not been assigned to a cluster).
- A similarity threshold, T_i. This is used to establish if two ants are sufficiently similar to be nestmates. This similarity threshold models the odor template learned by the ants during their youth and is estimated during a learning phase. It can also be adjusted during the meetings process.
- The age, A_i. This is in fact a counter which counts the number of meetings to which the ant has participated and is used in computing some mean values.
- An adaptive parameter, M_i. This measures the ant's perception of its nest's size and is initially set to 0.
- An adaptive parameter, M_i^+. This measures the ant's perception of the acceptance degree by the other members of its nest. As M_i^+ is larger as the ant is better integrated in its nest.

The clustering process consists of three main phases:

- *Threshold learning phase.* The aim of this phase is to estimate the value of T_i for each ant i. The similarities between an ant i and other randomly selected k_T ants are computed and the maximum ($\max\{S(i,\cdot)\}$) and the average ($\langle S(i,\cdot)\rangle$) of these similarities are determined. The estimation of T_i is $(\max\{S(i,\cdot)\} + \langle S(i,\cdot)\rangle)/2$. A particularity of the similarity measure used in AntClust is the fact that it is always in $[0,1]$. In the case of two n-dimensional numerical data, x_i and x_j, it is defined by:

$$S(i,j) = \frac{1}{n}\sum_{k=1}^{n}\left(1 - \frac{|x_i^k - x_j^k|}{|\max x^k - \min x^k|}\right) \quad (1)$$

- *Random meetings phase.* Random pairs, (i,j), of distinct ants are selected k_M times. For each pair, the similarity $S(i,j)$ is computed and is verified if these ants accept each other. Ants i and j accept each other if their similarity is larger than both thresholds: $S(i,j) > T_i$ and $S(i,j) > T_j$ (such an acceptance situation is denoted by $Accept(i,j) = True$). Depending on the acceptance relation and on the current labels L_i and L_j some behavioral rules are applied. The effect of these rules consists in possible modifications of labels and of parameters M_i and M_i^+.
- *Clusters refining phase.* In this stage the clusters having a small number of elements and low M^+ values are eliminated and their elements are assigned to other clusters based on their similarities and on the values of parameters M and M^+. The first version of clusters refining proposed in [Labroche, 2002] has been improved in [Labroche, 2004].

In the meeting phase the clustering process is controlled by the following behavioral rules applied to each meeting of two ants (i,j):

- *R1: new nest creation.* If $Accept(i,j) = True$ and $L_i = L_j = 0$ then $L_i := L_{max}+1, L_j := L_{max}+1$

where L_{max} is the maximal value of labels assigned up to the current step.

- *R2: including an ant into an existing nest.* If $Accept(i,j) = True$ and $L_i = 0$, $L_j \neq 0$ then $L_i := L_j$ (if $L_i \neq 0$, $L_j = 0$ then $L_j := L_i$).
- *R3: positive meeting between two nestmates.* If $Accept(i,j) = True$ and $L_i = L_j \neq 0$ then increase M_i, M_i^+, M_j and M_j^+.
- *R4: negative meeting between two nestmates.* If $Accept(i,j) = False$ and $L_i = L_j \neq 0$ then the ant having the smaller acceptance degree is eliminated from its nest (its label and its parameters M and M^+ are set to 0), the parameter M of the other ant is increased while the parameter M^+ is decreased.
- *R5: meeting between ants belonging to different nests.* If $Accept(i,j) = True$ and $L_i \neq L_j$, $L_i \neq 0$, $L_j \neq 0$ then the ant with the lower M is included in the nest of the other ant and M_i and M_j are decreased.

Increasing and decreasing the values of the parameters M and M^+ is based on the following relations:

$$\text{increase}(v) = (1-\alpha)v + \alpha \qquad \text{decrease}(v) = (1-\alpha)v \tag{2}$$

with $\alpha \in (0,1)$. These relations ensures that the values of M and M^+ are always in $[0,1)$. It is easy to prove that starting from a nonzero value v_0 the sequence of values generated by successive increasing steps converges to 1 while the sequence of values generated by successive decreasing steps converges to 0.

The general structure of the algorithm is presented in Algorithm 1.

Before analyzing the ability of Algorithm 1 to deal with noisy data we shall present some remarks on the parameters involved in the algorithm. The parameter α used in increasing and decreasing the parameters M and M^+ should have rather a small value (e.g. $\alpha = 0.1$ or $\alpha = 0.2$) otherwise it leads to a too fast increase or decrease of M and M^+. The parameters θ_r and θ_a are some threshold values. In all experiments we used $\theta_r = \theta_a = 0.1$.

The number k_T used to estimate the thresholds T_i should be high enough in order to lead to a reliable estimation. In experiments we used $k_T = 30$. We have to remark here that this separate learning phase can be replaced by a continuously adaptation of T_i during the meetings phase (of course the first meetings, when the thresholds are still undefined, don't define clusters but only adjust the values of T_i). No significant differences between the results obtained by these two variants have been noticed in our experiments.

An important parameter is the number of meetings, k_M. It has to be large enough in order to allow each ant to participate in several meetings. The minimal value of k_M should be the number of ants, m, but in order to estimate the parameters M and M^+ and allow clusters to reorganize it should be larger (mainly if the clusters are not very well separated). In [Labroche, 2002] is suggested a value proportional to m, $k_M = \kappa m$. In this case it is easy to prove that the averaged number of meetings each ant participates is 2κ (for a proof see the Appendix 1).

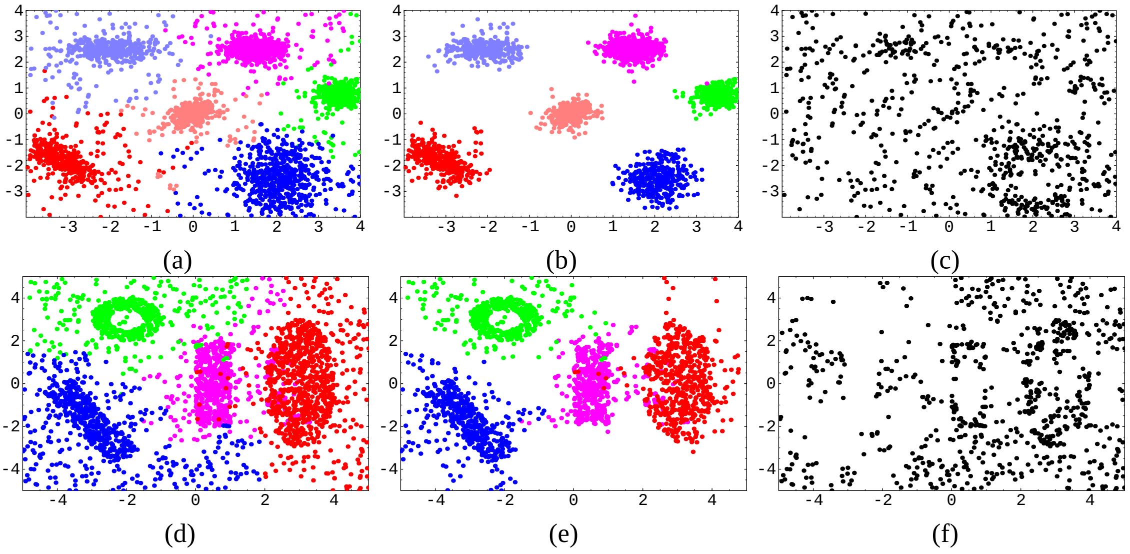
Algorithm 1 AntClust algorithm

```
for all $i \in \{1,\dots,m\}$ do
    $L_i := 0$; $A_i := 0$; $M_i := 0$; $M_i^+ := 0$;
end for
{Threshold learning:}
for all $i \in \{1,\dots,m\}$ do
    sample $k_T$ ants and compute
    $max\{S(i,\cdot)\}$ and $\langle S(i,\cdot)\rangle$;
    $T_i := (max\{S(i,\cdot)\} + \langle S(i,\cdot)\rangle)/2$;
end for
{Random meetings:}
for all $k \in \{1,\dots,k_M\}$ do
    Select a random pair $(i,j)$
    Increase the age: $A_i := A_i + 1$; $A_j := A_j + 1$;
    Compute $S(i,j)$
    Apply the rules R1-R5
end for
{Clusters refining:}
for all identified clusters do
    compute $N_{cluster}$ (the ratio of data belonging to the cluster) and
    $\langle M^+\rangle$ (averaged value of $M^+$ for all ants in the cluster)
    compute the acceptance probability
    $P_a = \alpha\langle M^+\rangle + (1-\alpha)N_{cluster}$
    if $P_a < \theta_r$ then
        remove the cluster (all its elements are reset)
    end if
end for
for all ants having $M^+ < \theta_a$ do
    assign the ant to the cluster of the most similar ant, $j$, which belongs to a nest ($L_j \neq 0$) and has a high enough acceptance degree ($M_j^+ > \theta_a$)
end for
```

We analyzed the behavior of AntClust for two synthetic noisy data sets. The first set consists of 6 ellipsoidal clusters, generated by using a bi-dimensional normal distribution, superposed with an uniform noise (see Fig. 1(a)). The clusters have different sizes, shapes and orientations (obtained by using different parameters of the normal distribution). The number of points in clusters is 2500 and the number of noisy points is 500.

The second set consists of 2050 points grouped in 4 clusters having different geometric shapes (the points have been uniformly generated in the interior of these geometric shapes) and 750 points uniformly distribute in the exterior of the geometric shapes, representing the noise (see Fig. 1(d)).

`AntClust` (applied with $k_M = m^2/10$ for the first data set and $k_M = m^2/2$ for the second data set) identified 6 and 4 clusters respectively as can be seen in Figs. 1(a) and 1(d). Since M^+ is a measure of the acceptance degree of an ant by its nestmates it seems natural to interpret it also as a level of its significance for the cluster. This means that data having low values for M^+ could be considered as candidates to be noise. The critical issue here is how to choose a threshold



(a) (b) (c)

(d) (e) (f)

Figure 1: (a),(d) Results obtained by applying the original `AntClust`; (b),(e) Clusters identified by ignoring the points for which the acceptance parameter (M^+) is lower than the quartile of values of M^+ for all data; (c),(f) The ignored points (estimation of the noise)

.

on M^+ values. Since all M^+ values are in $[0,1)$ an absolute value could be used (e.g. 0.1). On the other hand a relative value based on some order statistics could be also applied. The results obtained by using as threshold the quartile value of M^+s are illustrated in Figs. 1(b),(c) and 1(e),(f). In the case of ellipsoidal clusters the result is acceptable but in the case of geometrical clusters the noise is not very well identified. This suggested us to use also a measure related to the data density.

3 Introducing Density Information in `AntClust`

In order to introduce in `AntClust` density information we propose to attach to each ant, i, a new parameter, D_i. This parameter will contain an estimation of the ant's perception on the density of the region were it is placed. This parameter is set to zero at the beginning and at each meeting between the ant i and a different ant j it is adjusted by $D_i := D_i + \Delta_i$ where

$$\Delta_i = \begin{cases} \exp\left(-\frac{(1-S(i,j))^2}{2\sigma_i^2}\right) & \text{if } 1-S(i,j) \le \sigma_i \\ 0 & \text{if } 1-S(i,j) > \sigma_i \end{cases} \tag{3}$$

where σ_i are parameters controlling the influence area of each data. Usually $\sigma_i = \sigma$ for all i. This density is similar to that used in DENCLUE but instead of computing it by a systematic search of the neighborhood as in DENCLUE it is estimated based on the random meetings of ants. These computations do not significantly increase the complexity of `AntClust`.

After the meetings phase, the parameter D_i is divided by the ant's age, A_i, the density estimation being always in $[0,1)$. The first question concerning D_i is if it really offers different information than M_i^+. Besides the fact that it is differently computed than M^+, results on various data sets suggest that they are not necessary positively correlated (see Fig.2). Figure 2 also illustrates that while M^+ takes values over the entire range $[0,1)$, the density parameter takes values only on a restricted subrange of $[0,1)$. This can be explained by the relation used in computing D_i which ensures that the range of density values could have a size of at most $1-\exp(-1/2) = 0.3934$ (for a proof see Appendix 2).

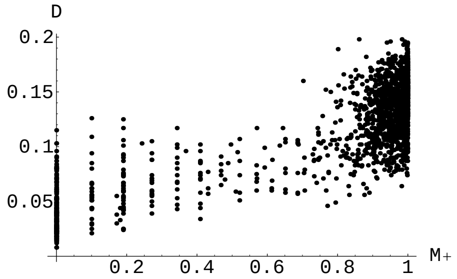


Figure 2: Dependence between the parameters M^+ and D for $\sigma = 0.1$

A first way of using the values of D_i is in the clusters refining phase: the elements without clusters are assigned to a cluster only if their density value is larger than the density's cuartile. Moreover, only the elements already belonging to clusters which have a density value larger than the density's cuartile are taken into account when we are searching for the most similar element. In this way a new class appear: that of unclassified data which can be considered noise.

The parameter σ plays an important role in identifying the clusters. We analyzed the influence of this parameter on the behavior of `AntClust` for the data set consisting of four geometric clusters (Fig. 1 (d)) and a noisy data uniformly distributed on the exterior of the geometric shapes. The algorithm should identify 5 classes, four corresponding

σ_i	No. of clusters		Classification error	
	mean	standard deviation	mean	standard deviation
0.1	5.7	0.48	0.137	0.009
0.25	5.5	0.52	0.153	0.007
0.5	5.7	0.67	0.153	0.005
0.75	5.7	0.67	0.148	0.007
$(1-T_i)/4$	5.4	0.51	0.152	0.006
$(1-T_i)/2$	5.8	0.62	0.158	0.011
$(1-\langle T\rangle)/4$	5.7	0.67	0.136	0.006
$(1-\langle T\rangle)/2$	5.6	0.51	0.136	0.008

Table 1: Influence of σ on identifying the four geometrical clusters and the noisy data cluster

to useful data and one to noisy data. In order to evaluate the clustering quality we computed an error measure introduced in [Labroche, 2002]:

$$Err = \frac{2}{m(m-1)} \sum_{i<j} \epsilon_{ij} \tag{4}$$

with

$$\epsilon_{ij} = \begin{cases} 0 & \text{if } c(i) = c(j) \text{ and } c'(i) = c'(j) \\ & \text{or } c(i) \neq c(j) \text{ and } c'(i) \neq c'(j) \\ 1 & \text{otherwise} \end{cases}$$

where $c(i)$ is the true cluster of the data corresponding to ant i and $c'(i)$ is the cluster to which the ant is assigned by the algorithm. Table 1 presents results concerning the number of identified clusters and the classification error for different values of σ. All these values are averaged over 10 independent runs of the algorithm. Besides some constant values of σ (0.1, 0.25, 0.5, 0.75) also values depending on the similarity thresholds of ants have been analyzed. Both individual parameters (e.g. $\sigma_i = (1-T_i)/2$) and determined by averaged values (e.g. $\sigma = (1-\langle T\rangle)/2$) were tried.

The results in Table 1 shows that the value of σ can be chosen depending on the values of the similarity threshold. The value $\sigma = (1-\langle T\rangle)/2$ proved to be adequate also for other test data.

On the other hand, the different roles which D_i and M_i^+ play suggest to use both of them in order to control the separation of useful data from noise. This means to split the data in four categories based on the values of their M^+ and D parameters. The separation is based on some threshold values T_M (threshold for the acceptance degree) and T_D (threshold for the density). The four categories are:

- *First category.* This contains all data i for which $M_i^+ \leq T_M$ and $D_i \leq T_D$ and corresponds to data having a high probability to be noisy (see Figs. 3(e) and 4(e)).
- *Second category.* This contains all data i for which $M_i^+ \leq T_M$ and $D_i > T_D$ and corresponds to data having a high estimation of the density but a low acceptance degree. These data couldn't be classified even if they belong to rather dense regions. Usually they are points at the border of clusters (see Figs. 3(f) and 4(f)) and are somewhat similar to border points identified in DBSCAN.
- *Third category.* This contains all data i for which $M_i^+ > T_M$ and $D_i \leq T_D$ and corresponds to data having a high acceptance degree in their cluster but a low estimation of the density. Usually these data belong to rather sparse regions but they participated to many meetings (see Figs. 3(c) and 4(c)). The existence of this category is explained by the different threshold used in estimation of M_i^+ (T_i) and in computation of D_i (σ).
- *Fourth category.* This contains all data i for which $M_i^+ > T_M$ and $D_i > T_D$ and corresponds to data which belong to dense regions and have been accepted by their clusters (see Figs. 3(d) and 4(d)).

σ_i	First category	Second category	Third category	Fourth category
0.1	2.1%	11%	1%	85.6%
0.25	0.6%	8.6%	5.4%	85.2%
0.5	1.9%	16.8%	4.1%	77.1%
0.75	0.5%	16.1%	4.6%	78.6%
$(1-T_i)/4$	16.6%	6.8%	19.6%	56.8%
$(1-T_i)/2$	17.3%	14%	14.6%	53.9%
$(1-\langle T\rangle)/4$	5.6%	17.8%	3.1%	73.3%
$(1-\langle T\rangle)/2$	2.9%	7.2%	1.4%	88.4%

Table 2: Distribution of data belonging to geometrical shapes in categories

The results presented in Figs. 3 and 4 have been obtained by using as thresholds, T_D and T_M the quartiles of parameters D and M^+ respectively computed over the entire set of data.

Some results concerning the distribution of data in these four categories are presented in Tables 2 and 3. Table 2 illustrates how the non-noisy data are distributed in the four categories for different ways of computing the parameters σ_i. As expected the largest percent of non-noisy data is assigned to the fourth category. Taking into account only the density information and ignoring the parameter M^+ the useful data would be represented by the union of the second and fourth category (this means that $D_i > T_D$).

On the other hand, Table 3 illustrates how are distributed the noisy data. In this case the largest percent of data are assigned to the first category and the union between the first and the third category represents an estimation of the noise. These results also suggests that an adequate choice of σ could be $(1-\langle T\rangle)/4$ or $(1-\langle T\rangle)/2$ (for these values the percents of non-noisy data categorized as useful data and of noisy data categorized as noise are larger than for other values). This means that the parameter σ could be chosen depending on the average of the similarity thresholds. Thus, besides the fact that introducing the density information doesn't significantly modify the complexity of `AntClust`, it neither introduces new parameters.

Besides the tests made on synthetic bi-dimensional data we also analyzed the relevance of density information in

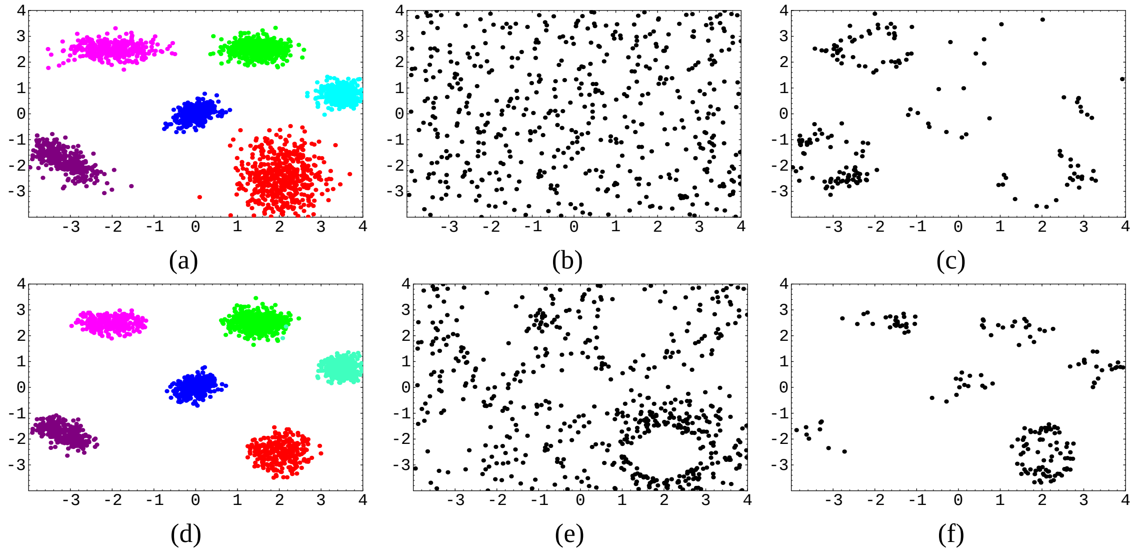


Figure 3: Ellipsoidal clusters. (a) the original clusters; (b) the uniform noise superposed on the clusters; (c) data in the third category; (d) identified clusters (fourth category); (e) identified noise (first category); (f) data in the second category

σ_i	First category	Second category	Third category	Fourth category
0.1	60.4%	4%	17.4%	18.2%
0.25	32.1%	12.4%	11.3%	44%
0.5	38.9%	18.1%	11.2%	31.8%
0.75	36.3%	16.9%	12.9%	33.9%
$(1-T_i)/4$	71.9%	2.2%	25.2%	2.7%
$(1-T_i)/2$	69.7%	3.6%	17.4%	9.3%
$(1-\langle T\rangle)/4$	73.4%	0.5%	23.2%	2.9%
$(1-\langle T\rangle)/2$	61.7%	1.3%	23.5%	13.5%

Table 3: Distribution of noisy data in categories

the case of other test data from UCI Machine Learning Repository, (www.ics.uci.edu/pub/ml-repos/machine-learning-database/). Tests for Iris database suggest that by taking into account the density information and identifying those four categories we can separate data which are difficult to classify. For instance, if the averaged percent of misclassified data in the entire data set is 9.46%, the fourth category contains 5.56% misclassified data.

4 Conclusions and open questions

The ability of an ant-based clustering algorithm (`AntClust`) in separating noise from data is analyzed. An analysis on the usefulness of both an existing parameter attached to ants (M^+) and that of a new parameter related to the density (D) is initiated. The computation of the density parameter and the postprocessing step of separating the data in different categories do not modify significantly the complexity of the algorithm. Moreover, since the parameter σ can be computed based on the similarity thresholds, it is not necessary to tune a new parameter.

The preliminary results are encouraging but a lot of things concerning the information carried by the parameters attached to ants still remains unrevealed. The values of thresholds, T_M and T_D used in postprocessing the data in order to identify the noise are mainly determined based on experiments and not on analytical reasons. Some theoretical results concerning the estimations of parameters M, M^+ and D computed during the meetings phase would be highly desirable.

Appendix 1

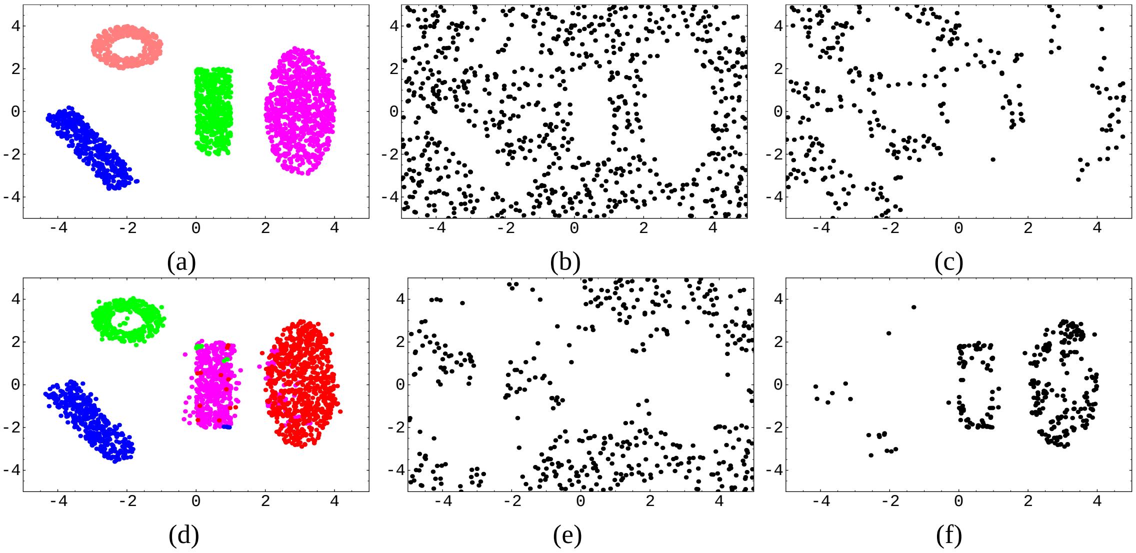
Proposition. *If $k_M = \kappa m$ then the averaged number of meetings for an ant is 2κ.*

Proof. Let Z_i be the random variable which corresponds to the number of selections of an ant i in the meetings phase (either on the first position or on the second position of a pair). In this analysis we suppose that the selection of both elements of a pair is uniform on $\{1,\ldots,m\}$. Thus Z_i has a binomial distribution, $B(N,p)$, of parameters $N = 2k_M$ and $p = 1/m$. Thus $P(Z_i = r) = C^r_{2k_M} p^r (1-p)^{2k_M - r}$, the mean value of Z_i is

$$\begin{aligned} E(Z_i) &= \textstyle\sum_{r=1}^{2k_M} r C^r_{2k_M} p^r (1-p)^{2k_M-r} \\ &= \textstyle\sum_{r=1}^{2k_M} r C^r_{2k_M} \frac{(m-1)^{2k_M-r}}{m^{2k_M}} \end{aligned} \tag{5}$$

For $k_M = \kappa m$ we obtain

$$\begin{aligned} E(Z_i) &= \textstyle\sum_{r=1}^{2\kappa m} r C^r_{2\kappa m} \frac{(m-1)^{2\kappa m - r}}{m^{2\kappa m}} \\ &= \frac{(m-1)^{2\kappa m}}{m^{2\kappa m}} \textstyle\sum_{r=1}^{2\kappa m} C^r_{2\kappa m} \frac{r}{(m-1)^r} \\ &= \frac{(m-1)^{2\kappa m}}{m^{2\kappa m}} \textstyle\sum_{r=0}^{2\kappa m-1} C^r_{2\kappa m-1} \frac{2\kappa m}{(m-1)^{r+1}} \\ &= \frac{(m-1)^{2\kappa m}}{m^{2\kappa m}} \cdot \frac{2\kappa m}{m-1} \cdot \frac{m^{2\kappa m-1}}{(m-1)^{2\kappa m-1}} = 2\kappa \end{aligned}$$



(a) (b) (c)

(d) (e) (f)

Figure 4: Geometrical clusters (a) the original clusters; (b) the noise superposed on the clusters; (c) data in the third category; (d) identified clusters (fourth category); (e) identified noise (first category); (f) data in the second category.

Appendix 2

Let us suppose that an ant i participates to k meetings, J is the set of indices for which $(1 - S(i,j)) \geq \sigma_i$ and card$J = k'$. Then D_i satisfies:

$$D_i = \frac{1}{k} \sum_{j \in J} \exp\left(- \frac{(1 - S(i,j))^2}{2\sigma_i^2} \right)$$

and since each term of the sum is between $\exp(-1/2)$ and 1 it follows that:

$$\frac{k'}{k} \exp(-1/2) \leq D_i \leq \frac{k'}{k}.$$

Since $k' \leq k$ it follows that the size of D_i range is at most $1 - \exp(-1/2)$.

Acknowledgments

This work was supported by the CNCSIS grant MindSoft/2005. The authors would like to thank anonymous referees for valuable comments and suggestions.

Bibliography

[Deneubourg, 1991] Deneubourg, J.L., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C., Chrétien, L.,(1991) ”The dynamics of collective sorting: robot-like ants and ant-like robots”, in J.A. Meyer and S. Wilson (Eds.), Proc. of the first Intern. Conf. on Simulation of Adaptive Behaviour: From Animals to Animats 1, MIT Press, Cambridge, pp. 356-365.

[Ester, 1996] Ester, M., Kriegel, H.P., Sander, J. and Xu, X., (1996) ”A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, Proc. of the Intern. Conf. on Knowledge Discovery and Data Mining, AAAI Press, pp. 226-231.

[Handl, 2003] Handl, J., Knowles, J. and Dorigo, M (2003) ”Strategies for the Increased Robustness of Ant-based Clustering”, Self-Organising Applications: Issues, challenges and trends, LNCS 2977, pp. 90-104.

[Hinneburg, 1998] Hinneburg, A. and Keim, D.A. (1998) ”An Efficient Approach to Clustering in Large Multimedia Databases with Noise”, Proc. of 4rd Intern. Conf. on Knowledge Discovery and Data Mining, AAAI Press, pp. 58-65.

[Labroche, 2002] Labroche, N., Monmarché, N. and Venturini, G. (2002) ”A New Clustering Algorithm Based on the Chemical Recognition System of Ants”, in Harmelen, F. van (Ed.) Proc. of the 15th European Conference on Artificial Intelligence, Lyon, France, pp. 345-349.

[Labroche, 2004] Labroche, N., Guinot, C. and Venturini, G. (2004) ”Fast Unsupervised Clustering with Artificial Ants”, X. Yao et al (Eds.): PPSN VIII, LNCS 3242, pp. 1143-1152.

[Lumer, 1994] Lumer, E., Faieta, B. (1994), ”Diversity and Adaptation in Populations of Clustering Ants”, in Proc. of the third Interm Conf. on Simulation of Adaptive Behavior: from Animals to Animats 3, MIT Press, Cambridge, MA, pp. 501-508.

[Nasraoui, 2004] Nasraoui O., Leon E., and Krishnapuram R. (2004), ”Unsupervised Niche Clustering: Discovering an Unknown Number of Clusters in Noisy Data Sets”, in A. Ghosh and L. C. Jain (Eds.), Evolutionary Computing in Data Mining, Springer Verlag, 2004.