Performance Evaluation of Ant Based Clustering

Ms.Vimal Gaur*
Kanika Khurana*
Girija*
Swati Chopra*

Abstract

Clustering analysis is an important function of data mining. Various clustering methods are need for different domains and applications. A clustering algorithm for data mining based on swarm intelligence called Ant-Cluster is analyzed in this paper. This paper shows the result of an experiment conducted to analyze this Ant Cluster Algorithm and evaluate its performance by generating a random test data. The results illuminate that Ant-Cluster algorithm can get clustering results effectively without giving the number of clusters and have better performance than k-means algorithm.

Keywords : Clustering Algorithm, Data Mining, Swarm Intelligence

1. Introduction

Clustering means the act of partitioning an unlabeled dataset into groups of similar objects. Each group, called a `cluster', consists of objects that are similar between themselves and dissimilar to objects of other groups. Clustering is the method by which like records are grouped together. Usually this is done to give the end user a high level view of what is going on in the database. From a machine learning perspective, clusters correspond to the hidden patterns in data, the search for clusters is a kind of unsupervised learning and the resulting system represents a data concept.

Major clustering methods are classified into five categories, i.e. partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. Some clustering algorithms integrate the ideas of several clustering methods. But each kind of clustering methods has its own limitation. For example, k-means algorithm, a widely used partitioning method, is sensitive to the initial objects that can induce local optimum, and needs user to specify the number of clusters. Swarm intelligence is a kind of evolutionary algorithm inspired by the behaviors of social animals. It has some advantages and characteristics, such as self-adaptation, self-government, and parallel computing, etc. It has been applied in

CSE Deptt. MSIT, C-4 Janakpuri, Affiliated to GGSIPU, New Delhi, India

the traveling salesman problem (TSP), quadratic assignment problem, graph coloring, job-shop scheduling, sequential ordering, and vehicle routing [2]. Ant based clustering is a heuristic clustering method that draws its inspiration from the behavior of ants in nature. Thus, the working of ant based clustering is quite different from those of ordinary clustering algorithms. While well known clustering algorithm like k-means gradually build or refine an explicit representation of a data set's partitioning, ant based clustering uses no such model but only implicitly generates the partitioning.

The rest of this paper is organized as follows. Section 2 introduces ant based clustering algorithms based on swarm intelligence briefly and illuminates the parameters and symbols used in this paper. Section 3 discusses the Ant-Cluster analysis and various accounting measures used in it. Section 4 do the comparison of Ant-Cluster algorithm with k-means algorithm. Finally, Section 5 concludes the paper and points out expectation for future research.

2. Ant-Based Clustering in the Literature

2.1 Ant-based clustering

Ant-based clustering and sorting was originally introduced for tasks in robotics by Deneubourg [4]. Lumer and Faieta [16] modified the algorithm to extend to numerical data analysis and it has subsequently been used for data-mining [17], graph-partitioning [15, 14, 13] and text mining [11, 20, 10].

2.2 Basics

The parameters and symbols used in this paper are illuminated as follows.

α swarm similarity coefficient;
r observing radius of each ant;
N the maximum of cycle times;
size : the size of the 2-dimension grid;
m_p the number of ants in each population;
p index of populations, p = 1, 2, 3;
p_p picking-up probability;
p_d dropping probability;
P_r random probability, $p_r \in [0, 1]$;
k_1 and k_2 threshold constants for computing p_p and p_d respectively;
ant_i the *i*th ant;
o_i the *i*th data object;

loaded and unloaded state of ant. If there is a data object on an ant, its state is **loaded** otherwise, its state is **unloaded**.

v_{high} : the speed of ants in high speed population;
v_{low} : the speed of ants in low speed population;
v_{max} : the miximal speed in variable speed population;
l : the maximum times of an ant moving with a same data object continuously.

2.3. Ant-Cluster algorithm

We proposed Ant-Cluster algorithm based on the existing research about ant-based clustering.

The high-level description of Ant-Cluster algorithm is shown in Algorithm.

ALGORITHM : A High-Level Description of Ant-Cluster

Initialization phase : Intitialize parameters (α, r, N, size, m_p, v_{high}, v_{low}, v_{max}, and *I*). Place data objects on a 2-dimension grid randomly, i.e. assign a pair of coordinates (x,y) to each data objects. Put three populations of ants with different speed on this 2-dimension grid. Initial state of each ant is **unloaded**.
While (cycle_time < N)
adjust α with specific step;
for (p = 1; p <= 3; p++)
for (*i* = 1; *i* <= m_p; *i*++)
if (**ant**$_i$ encounter a data object)
if (state of **ant**$_i$ is **unloaded**)

Compute the swarm similarity of the data object within a local region with radius ***r***, and compute picking-up probability p_p. Compare p_p with a random probability p_r. if $p_p > p_r$, ant_i pick up this data object, and the state of ant_i is changed to **loaded**.

else
if (state of ant_i is loaded)
if ant_i has already moved with the same data object *I* steps, the data object is dropped and the state of ant_i is changed to **unloaded** otherwise, compute the swarm similarity of the data object within a local region with radius *r*, and compute dropping probability p_d, Compare p_d with a random probability p_r, if $p_p > p_r$, ant_i drops this data object, and the state of ant_i changed to **unloaded**

end
end
end

3. Measures Accounting on Various Scales

3.1 Picking up and Dropping up probability

The various functions used to determine the location , speed and the corresponding behavior of ants are :-

The swarm similarity is computed by following formula :

$$f(oi) = \frac{1}{s} \sum_{o(j) \in Neigh(r)} \left[1 - \frac{d(o(i), o(j))}{\alpha} \right]$$

Where $f(o_i)$ is a measure of the average similarity of object o_i with the other objects o_j present in the neighborhood of o_i. S is the number of objects o_j. $d(o_i, o_j)$ is the distance between two objects o_i and o_j in the space of attributes measured with Euclidean distance.

The swarm similarity is transformed to picking-up probability p_p and dropping probability p_d by following formulas respectively.

$$p_p(o_i) = \left(\frac{k_1}{k_1 + f(o_i)} \right)^2$$

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

Wherein, k_1 and k_2 are two threshold constants assigned by users.

3.2. Correlation -based evaluation

The Pearson correlation coefficient provides information on the degree of linear relationship between the distributions of two variables. In the context of topographic mappings it can therefore be employed to determine the degree to which a mapping preserves a linear relationship between the distances in data-space (described by distribution X) and those in map-space (described by distribution Y):

$$P = \frac{\text{covariance}(X, Y)}{\text{Variance}(X) \bullet \text{Variance}(Y)}$$

P takes values in the interval [−1, 1], with 1 signifying perfect positive correlation,

−1 signifying perfect negative correlation, and 0 signifying a complete lack of linear correlation.

3.2.1 Inter-cluster Pearson correlation

When the real cluster distribution is known, the correlation of the respective distances between cluster centers in data-space and in map-space can additionally be computed. This gives insight into the preservation of global relationships, that is, the relative distances between clusters. We refer to this measure as the inter-cluster Pearson correlation.

3.2.2 Intra-cluster Pearson correlation

Similarly, the Pearson correlation can be applied to obtain more detailed information on the sorting within individual clusters. The correlation coefficient computed for the dissimilarities of the data items belonging to individual clusters provides information about the preservation of more local relationships. We refer to this measure as the intra-cluster Pearson correlation.

4. Performance Evaluation

In [LF94] the original algorithm was mainly evaluated using visual observation and by measuring spatial entropy.The quality of sorting was measured by the global_t and a dissimilarity measure.

As these do not reflect the global correlation well enough, [KSL98] used the overall Pearson Correlation (the degree of linear relationship between two variables) as an additional performance measure. Here, correlations of up to 0:65 were reported but not analyzed further. Importantly, it must be noted that the method is very sensitive to the choice of α and correlations like 0:65 are only achieved with the proper choice of α

This problem was already noted in [KSL 98], but no method was suggested how the proper α value could be found without resorting to manual experimentation. The proper choice of depends on the structure of the data and the solutio is to introduce an adaptive strategy.

To further analyze the performance, we measure Pearson correlation for the different no. of iterations The results are as shown below from our experiment

Figure1 shows the development of Pearson correlation on the test data set with speed and memory improvements averaged

No. of Iterations	Pearson Correlation
0	0.0015
50	0.19
100	0.33
150	0.45
200	0.52
250	0.55
300	0.58

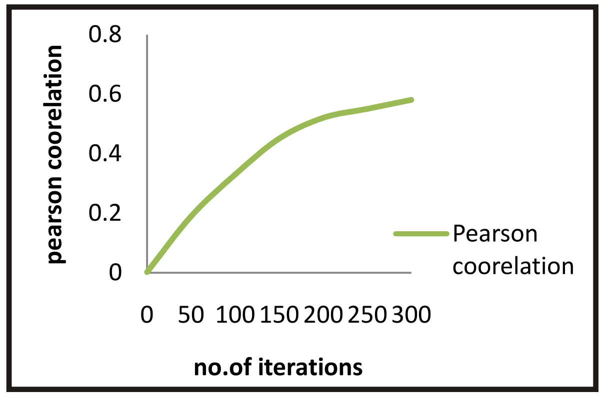


Figure 1

over 300 runs. It is clearly visible that the correlation remains very low. When analyzing the different correlation types further we can see that intra-cluster correlation and overall correlation both stay low.

As shown in figure 2 the performance of ant based clustering drops significantly, as it fails to reliably detect the four clusters. Clustering result is illuminated in figure 3

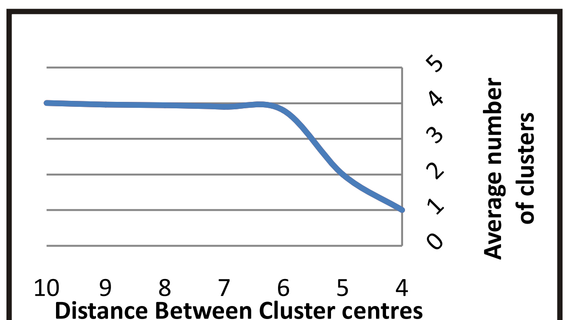
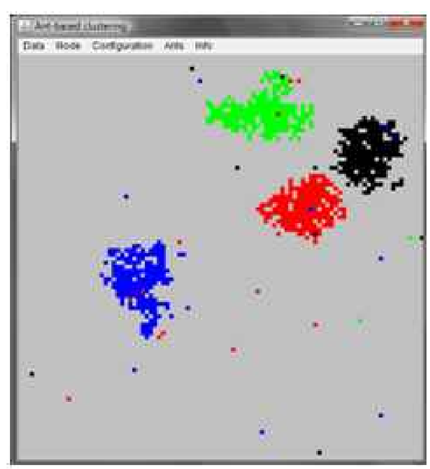


Figure 2



In figure 3, each cluster figures one dataset cluster. Objects in a cluster have some common characteristics and these characteristics can be obtained by comparing distribution of an attribute value in the whole data set with the one in a certain cluster.

5. Comparison over k-means Algorithm

5.1 k-means algorithm

The first algorithm we compare against is the well-known k-means algorithm. Starting from a random partitioning, the algorithm repeatedly (i) computes the current cluster center.(i.e .the average vector of each cluster in data space) and (ii) reassigns each data item to the cluster whose centre is closest to it. It terminates when no more reassignments take place. By this means, the intra-cluster variance, that is, the sum of squares of the differences between data items and their associated cluster centers, is locally minimized 'means' strength is its runtime, which is linear in the number of data elements, and its ease of implementation. However, the algorithm tends to get stuck in suboptimal solutions (dependent on the initial partitioning and the data ordering) and it works well only for spherically shaped clusters. It requires the number of clusters to be provided or to be determined (semi-) automatically. In our experiments, we run k-means using the correct cluster number. Random initialization is used and the best result out of 10 runs (in terms of minimum variance) is selected in order to reduce the effects of local optima.

5.2 Comparison with k-means algorithm

Ant based clustering is one of the few clustering algorithm that has the intrinsic capability to identify the number of clusters k in the data. Most other clustering methods (like the considered k-means) rely on the specification of k as an input parameter. This requires a priori knowledge or the interaction with another algorithm, which can be problematic, as current methods for the automatic determination of the number of clusters in a data set are rather limited On the other hand, the impossibility to specify k can also be considered a disadvantage of ant-based clustering and sorting. In specific applications the user might have precise ideas about the number of clusters to be identified, and in ant-based clustering and sorting, there is currently no possibility to adjust the number of clusters that are to be generated.

Secondly, it is the only algorithm to perform consistently well on all synthetic data sets, whereas k-means, have their individual problems. On the other hand, the results obtained on real data also show the limitations of ant-based clustering and sorting: if cluster structures on several levels exist, it only identifies the upper level ones. A recursive application of the algorithm on the resulting clusters might be a way, to be tested in future work, to overcome the problem.

Runtime

For the small data collections, the ant algorithm's runtimes are considerably higher than those of k-means, hierarchical clustering and 1D-SOM. However, it is worth observing that they scale linearly, such that ant-based clustering already outperforms the agglomerative scheme on the Yeast data set. As both k-means and 1D-SOM are affected by the rise in dimensionality, ant-based clustering even becomes the fastest method on the $ 10D-2D data.

6. Conclusion

We have re-examined ant-based clustering methods in the context of a real-life application and demonstrated the performance of the algorithms In contrast to previous studies on ant-based clustering, we have analyzed the clustering, in order to obtain an improved understanding of its performance relative to other methods for clustering. Ant-based clustering performs well in our comparison to the popular methods of K-means. In addition to that, the algorithm has a number of features that make it an interesting candidate for cluster analysis.

Firstly, there is its linear scaling behavior, which is attractive for use on large data sets as are frequently encountered today, e.g., in information retrieval. Also, the nature of the algorithm makes it fairly robust to the effects of outliers within the data. In addition, ant-based clustering has the capacity to work with any kind of data that can be described in terms of symmetric dissimilarities, and it imposes no assumption on the shape of the clusters it works with. Finally, an important strength of the algorithm is its ability to automatically determine the number of clusters within the data.

7. References

1. *P. Berthing. Survey of clustering data mining techniques. Technical report, Accrue Software, SanJose,California, 2002.*
2. *C. Blake and C. Merz. UCI Machine Learning Repository. Department of Computer Science, University of California.*
3. *E. Bonabeau, M. Dorigo, and G. Theraulaz. Swarm Intelligence – From Natural to Artificial Systems. Oxford University Press, New York, NY, 1999.*
4. *Cheng-Fa Tsai, Han-Chang Wu, Chun-Wei Tsai, "A new data clustering approach for data mining in large databases", Proceedings of the International Symposium on Parallel Architectures, Algorithms and Networks, Metro Manila, Philippines, pp. 22-24, 2002.*
5. *H.Azzag; N.Monmarché; M.Slimane; C.Guinot; G.Venturini, "AntTree: a new model for clustering with artificial ants", Proceedings of the 7th European Conference on Artificial Life, Dortmund, Germany,*
6. *Wu bin, Shi Zhongzhi, "A clustering algorithm based on swarm intelligence",Proceedings of the International Conferences on Info-tech and Info-net, Beijing, China, pp. 58-66, 2001.*
7. *Wu Bin, Zheng Yi, Liu Shaohui, Shi Zhongzhi, "CSIM: a document clustering algorithm based on swarm intelligence", Proceedings of the Congress on Computational Intelligence, Hawaiian, USA, pp. 477-482, 2002.*
8. *Yan Yang, Mohamed Kamel, "Clustering ensemble using swarm intelligence", Proceedings of the 2003 IEEE Swarm Intelligence Symposium, Piscataway, NJ, USA, pp. 65-71, 2003.*
9. *Handl, J., Knowles, J. and Dorigo, M, "Ant-based clustering: a comparative study of its relative performance with respect to k-means, average link and 1D-som", Technical Report TR/IRIDIA/2003-24. IRIDIA, Universite Libre de Bruxelles, Belgium, 2003.*
10. *Peng Yuqing, Hou Xiangdan, Liu Shang, "The K-means clustering algorithm based on density and ant colony", Proceedings of the International Conference on Neural Networks and Signal Processing, Nanjing, China.*