Winter 12-5-2004

An Ant-based Clustering Algorithm in Data Mining

Yong Tang

Yongkai Ma

An Ant-based Clustering Algorithm in Data Mining

Yong Tang, Yongkai Ma

Management School, University of Electronic and Science of China, Chengdu 610054, China
{TangYong, Mayongkai}@uestc.edu.cn

ABSTRACT

A new algorithm ABC (Ant based Clustering) inspired from behavior of the real ants is proposed for clustering in data mining. ABC employs some ants clustering in the searching space. Different from some clustering algorithms, instead of picking or dropping data objects, artificial ants in ABC find and merge the similar data points in their vision range, also a following and randomly route choosing strategy is introduced. Computational simulation results show that ABC is efficient, and the clustering results can be achieved in short time and maintain stably.

Keywords: data mining, clustering, ant colony algorithm

1. INTRODUCTION

Clustering is an important area in data mining. The purpose of clustering is to cluster a great deal of objects into groups according to the rule that is "things of one kind come together". The similarity of objects in a cluster should be as big as possible while the difference of objects in different clusters should be as big as possible too. Clustering has been applied in many domains, such as spatial data analysis, image processing, marketing and pattern recognition etc. Wildly used clustering algorithms are K-means, density based DBSCAN and network based STING etc [1,2]. Our ant based clustering algorithm, which is inspired by the collective behavior of real ants can avoid some shortages like (1) There is no need to determine the number of clusters artificially in advance in ABC. However, in general algorithms, users have to set the number as an input parameter. This requires rigorous domain knowledge for users, and on the other hand, the final clustering result is every sensitive to the initial number of clusters set in advance. But the truth is that, the number of clusters is not obvious especially when the data point is in a high-dimension space or the cluster rule is hard to tell. While in ABC, the clustering is going automatically, namely ABC might find the number of clusters by itself and when the clustering standard is chosen properly, the number of clusters and the results are credible and stable. (2) ABC can find clusters in any shape. Clustering algorithms based on the Euclidean distance output sphere-shape clusters, which is not a good way to describe the spatial distribution of data points in the space. In ABC, we introduce a clustering space or can be called as searching space. The position of a data point in the clustering space has no relations to its position in the high-dimensions data space. We randomly project data points in the data space into the clustering space, after doing that, we cluster them and the final distribution is our output, which indirectly indicates the distribution of data points in the original data space. (3) ABC can achieve the global optimization. Some algorithms are easily to stick into local optimizations. The ant based clustering algorithm proposed in our research is a typical nature-inspired bionic algorithm. Artificial ants in the algorithm simulate the clustering behaviors of real ants to do random searches in clustering spaces, and taken the data points as the objects needed to be clustered. Every ant chooses a similar object in its vision range to move, meanwhile update the mount of the pheromone on the route to guide other ants. Repeatedly, the similar data points are clustered, when the count of iterations is reach an initial number or the error is less than an initial value we output the result. This algorithm does not require domain knowledge, and is able to generate the global clusters automatically. (4) ABC can process large mount of data in batches. General algorithms need to process all the data in one time, which is unrealistic when the data scale is large for the space and time limitations. Because ABC is a self-heuristic algorithm, it can process new inputs based on present clustering results, and adjust the clustering results dynamically.

2. A DESCRIPTION OF CLUSTERING PROBLEMS IN DATA MINING

First, we give a brief describe of the clustering problems. Data collected from different domains are usually in high-dimensions. And the distribution is what clustering concerns. Mathematically, we have n discrete data points $X_1,\cdots,X_n$, which scattering in the d dimensions space $\Re$ and for every point $X_i=(X_{i,1},\cdots,X_{i,d})^T$. Then we can define the distance between any two data points X_i and X_j as $d_{X_i,X_j}=(\sum_{k=1}^{d} p_k(X_{i,k}-X_{j,k})^2)^{\frac{1}{2}}$, where p_k is the weight for the k dimension of data point X_i indicating the contribution of the k dimension under a certain of evaluating standard. To measure the result of the clustering also to see if the satisfaction of the clustering is met or not, error ε is introduced. For example, in the K-Means algorithm, after getting K

clusters, we can calculate the geometrical centers of these K clusters as $C_i, i=1,\cdots,K$ and their corresponding errors $\varepsilon=\sum_{j=1}^{K}\sum_{i=1}^{N_j} d_{X_i,C_j}$, where N_j is the number of data points fall into the same cluster whose center is C_j. When the error ε is smaller than a set value, we accept the clustering results otherwise we adjust the clusters. Existing clustering algorithms includes Hierarchical methods, Partitioning Methods, Grid-Based Methods and Density-Based Methods. And there are still some problems should be considered: (1) the outliers (2) applying users' domain knowledge in the clustering analysis (3) clustering large mount of data.

3. AN ANALYSIS OF CLUSTERING ALGORITHM

Ant in natural world is kind of typical social insect. By observing the collection behaviors of some objects such as food and corpses of ant colonies, biologists find that even without a centralized control mechanism, a colony of ants can do collections together in a very efficient way. This inspires us to deal our data points by simulating the collective behaviors of ant colony. Deneubourg [3] proposed a model to describe the clustering behaviors of ants, the basic idea is that ant can pick up and drop down objects with a probability rule as $P_{pickup}=(k_1/(k_1+f))^2$ and $P_{drop}=(f/(k_2+f))^2$ for picking up and dropping down respectively. And f is the coefficient describes an ant's sensitive ability of its environment, k_1 and k_2 are limitations. Lumer and Faieta [4] improved this model by using $d_{O_i,O_j}=(\sum_{k=1}^{d}(O_{i,k}-O_{j,k})^2)^{\frac{1}{2}}$ to denote the Euclidean distance between two objects O_i and O_j. And the environmental parameter of object O_i in a range of s is defined as $f(O_i)=\frac{1}{s^2}\sum_{O_j\in Neighbour_{s\times s}(i)}[1-d(O_i,O_j)/\alpha]$. We see that if object O_i is similar to its environment, $f(O_i)$ will be relatively big, otherwise $f(O_i)$ will be small. Thus the probability for picking up and dropping down is redefined as: $P_{pickup}(O_i)=(k_1/(k_1+f(O_i)))^2$ and $P_{drop}(O_i)=\begin{cases}2f(O_i) & if\ \ f(O_i)<k_2\\ 1 & if\ \ f(O_i)\ge k_2\end{cases}$, again k_1 and k_2 are limitations. Kuntz [5,6,7] applied this method in VLSI and Wu [8] did related work in the clustering of WEB document. In the model of Deneubourg [3], Lumer and Faieta [4], at first, data points are randomly projected onto a 2-dimension plane, notice that the positions of data points on the plane have nothing to do with their positions in the data space. Then artificial ants do random movements, they pick up a data objects with probability P_{pickup} and drop data objects with P_{drop}. After some runs, the algorithm output the clustering results. The potential problem this kind of algorithm may have is that it might get stick in a halfway and fail to get data points clustered thoroughly, in other words, similar data points which should be clustered in a group, form several sub-clusters in the space and cannot be clustered together in a cluster any more. The reason for this is that $f(O_i)$ is used to denote the similarity of O_i to its local environment, when O_i is surrendered by some similar data points, the probability for O_i to be picked up will be small, consequently O_i might not be picked up any more, which means the O_i is fixed on the 2-dimension plane, its position will not be changed. Essentially, this result is locally optimized rather than global optimization we are seeking for. Monmarche [9] combined this kind of algorithm with K-means to achieve a fast clustering speed. Ant based clustering uses lots of ants searching in the space randomly with an enforcement of the pheromone on their routes to guide other ants. This positive feedback able the algorithm to guild further searches based on the already available searching experience and knowledge. The original model is first proposed by Dorigo [10] and performs very well in solving TSP and JSP, etc [11]. Yang [12] defines the mount of the pheromone between objects O_i and O_j as $\tau_{O_i,O_j}(t)=\begin{cases}1, d(O_i,O_j)\le r\\ 0, d(O_i,O_j)>r\end{cases}$, and O_i is clustered into O_j with a probability $P_{ij}=\tau_{ij}^{\alpha}\eta_{ij}^{\beta}\Big/\sum_{s\in Neighbour(j)}\tau_{sj}^{\alpha}\eta_{sj}^{\beta}$, where τ_{ij} denotes the mount of pheromone, η_{ij} is the knowledge of the environment, which can be defined as $\eta_{ij}=1/d(O_i,O_j)$. Obviously, the smaller the distance between O_i and O_j, the more the mount of the pheromone, namely the higher the probability O_i and O_j been clustered in a same cluster.

4. NEW ANT BASED CLUSTERING ALGORITHM, THE IDEA AND STEPS

In our research we propose a new ant based clustering

algorithm. With a closer look at the behavior of real ants we believe that the vision ability of ants is limited in a certain range, namely their eyesight is limited, we use r to present the range ants can see. And ants are different, so they should not be treated in a same way. In our algorithm, some ants are big and some are small, namely their range of eyesight is in a direct proportion to the number of data points the $Cell_i$ has where the ant occupies, then we get $r = k\sqrt{\frac{l^2}{N}}\Big/2, k = 1,2,\cdots,$ k for the number of data points been clustered on $Cell_i$, l is the width of the plane, N is the number of ants. So, an ant will see farther if $Cell_i$ has more data points. We let bigger ants search first in our algorithm, which is described as following:

(1) Randomly projects data points onto a 2D plane. One cell can only have one data point at most.

(2) Randomly place m ants, every ant carries the data points on a $Cell_i$.

(2.1) For ant Ant_i, we calculate its eyesight range $r = k\sqrt{\frac{l^2}{N}}\Big/2, k = 1,2,\cdots,$ where k is the number of data points been clustered on $Cell_i$.

(2.2) Calculate the average center of the data points on $Cell_i$ where ant Ant_i is. The center is $\overline{C_i} = \frac{1}{k}\sum_{j=1}^{k} O_j$. If ant Ant_i finds there are data points on cell $Cell_j$, and $d(\overline{C_i}, \overline{C_j}) \le d$, where d is a gate-value, then $Cell_j$ is added into a candidates set $Candidate$. After the search, ant Ant_i finds out the most similar cell $Cell^*$ then merges all the data points carried by Ant_i with the data points on $Cell^*$ and at the same time, we update the mount of pheromone on the path between $Cell_i$ and $Cell^*$ to 1, namely $\tau_{Cell_i, Cell^*} = 1$ which indicates that this path has been walked by Ant. Ant Ant_i is always trying to follow the path, which has been taken by those ants set off from $Cell_i$, in other words, it attempts to find $Cell_k$, where $\tau_{Cell_i, Cell_k} = 1$. If this attempt is failed, we randomly move the data points of cell $Cell_i$ to a cell $Cell_r$ on the plane, to avoid mistakes we must make sure that there are no data points on $Cell_r$.

(3) When all ants finish their moves, we calculate the total error $\varepsilon = \sum_{j=1}^{K}\sum_{i=1}^{N_j} d_{X_i, C_j}$, once final condition is met, namely $\varepsilon \le \varepsilon_0$, we output the clustering results and stop, otherwise go back to 2 and continue. Of course, we can set a count number $Count$, when the iterations gets $Count$ or there is no changes in the clustering results, we stop and output.

In (1), as other algorithms do, we randomly project data points onto the 2D plane, we must remember that the 2D plane is a search space, the positions of n-D data points in the data space have nothing to do with the positions in the 2D plane. In this algorithm, artificial ants can handle both single data points and data points sets as well. Thus the problem we mentioned before can be avoided, that is a single data point cannot be moved for the similarity to its environment. By introducing the pheromone, we make the system have a kind of memory so that former searching experience can be reused to guild following searches. Artificial ants have vision range and in the searching process adapting a route-following and random choosing strategy make them closer to real ants in the nature.

5. SIMULATIONS

We apply the discussed ant based clustering algorithm in simulations. In this research, we use IRIS [13] in UCI machine learning database as our test dataset. There are 150 data points in the dataset and every data point has four parameters. There are 3 clusters: Iris Setosa, Iris Versicolor and Iris Virginica, each cluster has 50 data points. Our simulation environment is: OS: Windows XP, CPU: P4 1.5GHz, Memory: 128MB, Language is C. Our settings are: 10 ants, width for the plane is 15, $Count$ is 100, $Count$ will descend by one. After multiple simulations, we find that the ABC is good: average running time is less than 10 seconds, the number of clusters found by ABC is 3, mistaken points is between 4 and 5. One of the best run, only one data point is mistakenly clustered that is the NO.124 data point should be clustered in Iris Virginica but is clustered in Iris Versicolor instead. Wu did similar work [8].

Fig.1- Fig.4 is the process in a run. Fig.5 is the process of a run. We see that when $Count = 25$, we get 3 clusters and the result maintain stably.

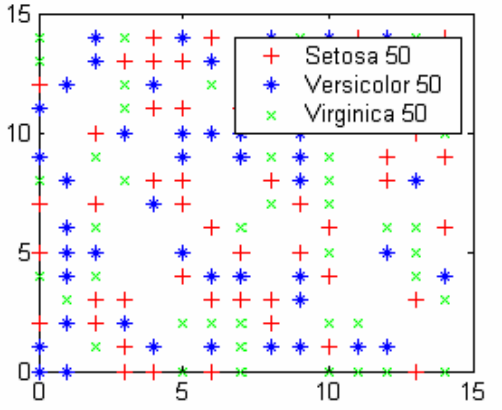

Fig.1 Count=0

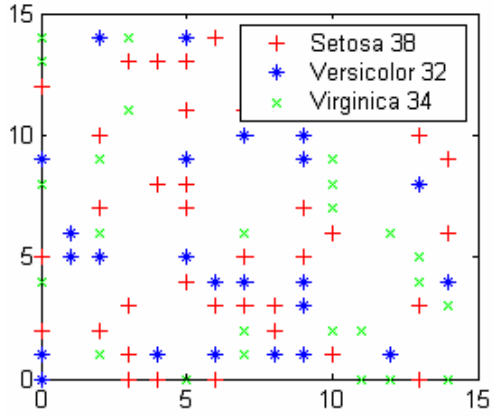

Fig.2 Count=5

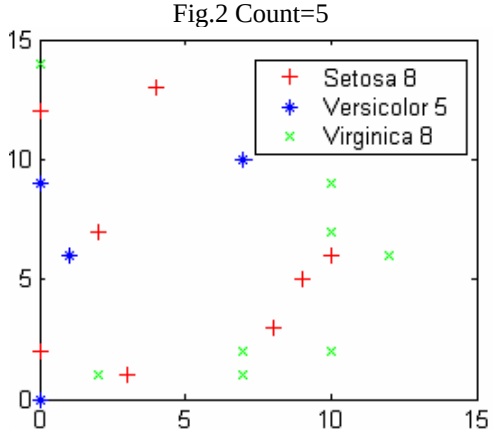

Fig.3 Count=18

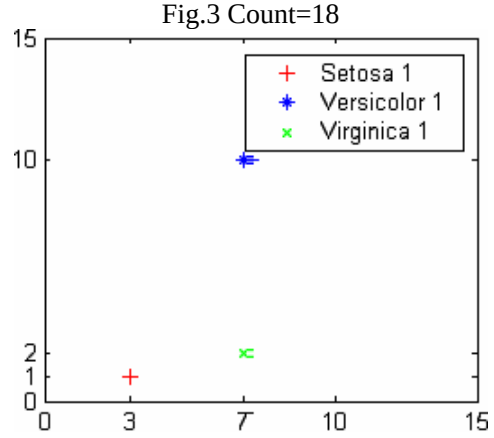

Fig.4 Count=25

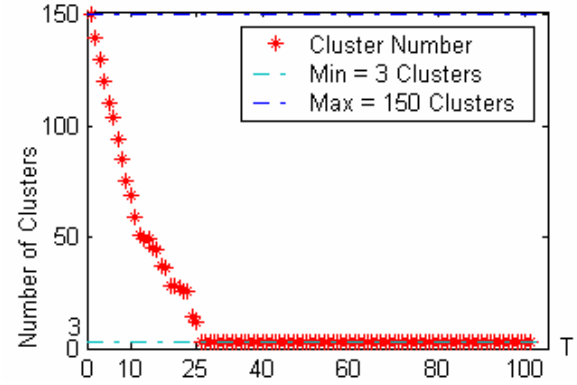

Fig.5 A process of clustering

6. CONCLUSION

Clustering analysis is one of the major areas in data mining and it has been widely applied in many domains. Clustering problems are usually in high dimensions and large scale. Nature-inspired algorithms have been used in many domains. We proposed an ant-based algorithm applying artificial ants in the searching space. Vision range concept is introduced with following strategy and random choosing strategy. The simulations show that the algorithm is fast and able to find the number of clusters automatically. As further work, we can consider: (1) At first we use ABC to find the number of clusters automatically, then we use other algorithms to do further clustering. (2) In ABC, we introduce a less similar data points kicking strategy to assure the correctness and reliability.

Applying nature inspired algorithms like ant colony algorithm to complex problems such as multiple objectives optimization, general function optimization and clustering analysis is something of swarm intelligence. As we have mentioned before, the original ant colony algorithm can solve every complex combinatorial optimization problems especially those can be abstracted into network optimization problems such as TSP etc. Swarm intelligence has a very important feature that is despite of the lacking of a centralized controlling mechanism or information spreading ways in certain structural forms ant colony can do things very effectively. Economic and social systems on complex networks can learn a lot from the collective behaviors of ant colony. The route choosing strategy used in the multiple objectives optimization inspires us to consider the decentralized collective behaviors in a brand new way. In other words, the artificial ant colony is a typical multiple agents system and the ant based clustering algorithm will be a good choice in the network clustering.

REFERENCES

[1] Han J., Kamber M. *Data Mining: Concepts and Techniques*, Beijing: Higher Education Press, 2001.

[2] Jain A. K., Murty M. N., Flynn P. J., "Data Clustering: A Review". *ACM Computing Surveys*,

Vol. 31, No. 3, pp264-323, 1999.

[3] Deneubourg J. L, Goss S., Franks N. et al., "The dynamics of collective sorting: Robot-like ants and ant-like robots", *proceeding of the First International Conference on Simulation of Adaptive Behavior: From Animals to Animats*. MIT Press, Combridge, MA, pp356-363, 1991.

[4] Lumer E. D., Faieta B., "Diversity and Adaptation in Populations of Clustering Ants" *proceedings of the third International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, MIT Press, Combridge, MA, pp501-508,1994.

[5] Kuntz P., Snyers D., "Emergent Colonization and graph partitioning" *proceedings of the third International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, MIT Press, Combridge, MA, pp494-500, 1994.

[6] Kuntz P., Layzell P., Snyers D., "A colony of Ant-like agents for partitioning in VLSI technology", *proceedings of the fourth European Conference on Artificial Life*, MIT Press, Combridge, MA, pp417-424, 1997.

[7] Kuntz P., Snyers D., Layzell P., "Astochastic heuristic for visualizaing graph clusters in a bi-dimernsional space prior to partitioning", *Journal of Heuristics*, No. 5, pp327-351, 1999.

[8] Bin W., Zhongzhi S., "A clustering algorithm base on swarm intelligence", *proceedings of IEEE 2001 International Conferences on Info-tech & Info-net*, Beijing, China, pp58-66, 2001.

[9] Monmarche N., "On data clustering with artificial ants", *Data Mining with Evolutionary Algorithms: Research Directions AAAI Workshop*, AAAI Press, pp23-26, 1999.

[10] Dorigo M., Maniezzo V., Colomi A., "The ant system: optimization by a colony of cooperating agents" *IEEE Trans. System man Cybernet. B*, Vol. 26, No. 1, pp29-41, 1996.

[11] Dorigo M., Caro G. D., Stutzle T., "Ant algorithms", *Future generation computer systems*, Vol. 16, No. 8, ppv-vii, 2000.

[12] Xinbin Y., Jinggao S, Dao H., "A New Clustering Method Based on Ant Colony Algorithm" *proceedings of the 4th World Congress on Intellegent Control and Automation*, Shanghai, China, pp2222-2226, 2002.

[13] Fisher R. A., "The Use of Multiple Measurements in Axonomic Problems" *Annals of Eugenics*, 7, pp179-188, 1936.