The Ant Colony Optimization of Clustering Problems in Data Mining

To cite this article: Xu Shijie 2018 J. Phys.: Conf. Ser. 1087 032023

View the article online for updates and enhancements.
The Ant Colony Optimization of Clustering Problems in Data Mining

Xu Shijie
Shandong Labor Vocational Technical College, Jinan 250022, China
19697672@qq.com

Abstract. Clustering analysis is widely used in engineering, it is an important part of data mining. Ant colony algorithm has robust and adaptable characters as new optimization algorithm. This research describes ant colony clustering algorithms characteristics, according to the existing ant colony algorithms. Although ant colony algorithm has obtained many improvements, there are many researches and applications in data clustering, but there are still some deficiencies that need further study.

1. Introduction
Cluster analysis is an important area in the search of data mining [1]. It is an effective means for people to understand and explore the internal relationship between things. The traditional clustering algorithms are mainly divided into five categories: division method, hierarchical method, density-based method, grid-based method and model-based method[1]. Inspired by the biological evolution mechanism, scientists have proposed many new methods to solve complex optimization problems. In 1991 the Italian scholar A. Dorigo et al. proposed an ant colony algorithm, which is a new optimization method [2]. Subsequently, he and other scholars proposed a series of algorithms concerning ant colony and applied it to the solution of complex combinatorial optimization problems, such as Traveling Salesman Problem (TSP) and scheduling problems[3]. The basic principles of these algorithms are easy to understand and have been applied to circuit design, text mining and other fields. This paper discusses in detail the basic principles and performance of the existing ant colony algorithm cluster analysis, and puts forward the need for improvement on the basis of summarizing and summarizing, in order to promote the application of ant colony algorithm in a wider area.

2. Ant Colony Clustering Algorithm Classification
In general, the ant colony algorithm-based clustering method can be divided into several types in principle: (1) using the principle of ant foraging, using pheromone to achieve clustering; (2) using ant self-aggregation behavior clustering; (3) heap formation data clustering; (4) Clustering using ant chemical recognition system[4].

2.1. clustering on ant foraging
The ants' foraging-based clustering algorithm The ant's foraging process can be divided into searching for food and carrying food. Each ant releases pheromones in its path during exercise and is able to perceive pheromones and their intensities. The basic idea of cluster analysis based on traces of ant pheromone is as follows: Considering data as ants with different attributes, the clustering center is the
“food source” that ants are looking for, and then the data clustering process can be regarded as an ant searching for food sources. process.

2.2. Clustering of Ant Self-aggregation Behavior

Clustering algorithm based on ant self-aggregation behavior. Ants can construct a tree structure through self-aggregation behavior, which is called AntTree. The data is represented by ants and represents the nodes of the tree. Initially, the ant places a fixed point called a fulcrum, which is equivalent to the root of the tree. The ants move on this tree or ants that are already fixed on the tree to find a suitable position for them. It is assumed that the ant can reach anywhere in the tree and stick to any position in the structure, but in the process of formation of the structure tree, the ant tends to be fixed at the end of the tree branch [5]. The similarity between the local structure of the tree and the data represented by the ant guides its movement. When all the ants are fixed in the tree, the algorithm ends and the dataset is divided.

2.3. Ant heap formation principle of clustering

The basic mechanism for clustering ant heaps based on the principle of ant heap formation is the process of worker ant accumulation of ant corpses. The small ants continue to attract worker ants to accumulate more dead ants, and the ant piles gradually increase due to positive feedback.

2.3.1 Algorithm Based on Agent Model

In 1991, Deneubourg[6] proposed an agent-based model to simulate the ant behavior to cluster data, and artificial ants moved along the grid unit. A data object will pick it up with a probability when it hits the object. The ant carrying the object. When the object to an empty cell or object is similar to an adjacent object, it is dropped with a certain probability. The probability of dropping also depends on the evaluation of the density of surrounding object types.

2.3.2 Ant-Based

Clustering is actually an improvement to the above algorithm. The algorithm mainly modified several important features, first modify the local similarity density. Second, use a short-term register to note the position of the previous drop object, when picking up an object can refer to the register to guide it to move, so that can be to the nearest Lowering the position of similar objects moves to reduce time complexity. Again, consider increasing the perceived radius to increase the neighborhood when computing, and combine some small clusters into one large cluster. Finally, make appropriate modifications to the dissimilarity factor.

2.3.3 Hybrid Clustering Methods

Although the neighborhood density function has been modified above, the number of final clusters is high and convergence is slow. So in 2000 Monmarche suggested that a unit could contain multiple objects, and each unit that owns an item would be a cluster. Later, Monmarche suggested clustering with the K-means method. The main process is as follows: Initially, the initial cluster is also formed using the ant algorithm. Since the division of time is too long, so before the algorithm converges on the termination of the algorithm, resulting in errors created division division, so the use of k-means algorithm to remove the small classification errors and assign "free" object, that there is still time to stop the algorithm alone. There is an object on the unit or an object that the ant is still carrying. Although this can remove the errors in the classification, the k-means algorithm is a local optimization algorithm and cannot obtain high-quality clusters, so it is necessary to use the ant-based algorithm again. This time it is the application of algorithms on the object heap rather than on individual objects. The previous ant-based algorithm also applies to the heap, which can be picked up or dropped again as a single object, again forming a new cluster. However, there are still unallocated heaps as before, so the final partitioning is again done using the k-means algorithm. This time, because the input provided to k-means is already very close to the best, the output is of high quality.
2.4. Clustering Using Ant Chemical Identification System

Clustering method based on ant chemical recognition system In order to protect one's nest from attack by the enemy or diners, the ants in reality must have the ability to distinguish between partners and enemies. When the two ants meet each other, the odors emitted by the other's epidermis (also called labels) are checked and compared with their own templates. Templates are acquired by ants during their early years and are continually updated as they grow. Tags are determined by the constant exchange of chemicals between ant genes and ants. Companions establish a group odor by constantly exchanging chemicals. groups have different odors and the same group share the same odor is the so-called “group circle” and is also the basic principle of the chemical identification system. In 2002, Labroche et al[7], proposed a clustering method called Ant Cluster, which is mainly based on the principle of chemical recognition system.

The above simple analysis of the popular clustering algorithm of ant colony, for each method from its basic ideas, clustering principles and the main steps are discussed and analyzed. Their difference lies in the difference in the communication medium between the individual ants, and some of them completely avoid the smell according to the odor. According to the odor communication, it is divided into two types: (1) according to the pheromone left on its path, such as the clustering method based on ant foraging; (2) based on the scent carried by the ant itself, such as based on ant chemical recognition Systematic clustering method. In addition, according to the spatial distribution of the object, the interaction of ants is guided to complete the clustering. For example, the clustering method based on the formation principle of the ant heap is based on the distribution of objects on the grid; The ant colony clustering method has many unique characteristics, such as flexibility, robustness, distribution and self-organization, etc. These characteristics make it very suitable for the problem of solving the problems of distribution, dynamics, and interleaving in essence[8].

3. Ant colony algorithm clustering application

The vocational school campus structure model from the wisdom of vocational colleges and universities hardware infrastructure, public support platform, intelligent software applications, integrated information services, scientific decision support, security protection, operation and maintenance services management, uniform standards from the bottom up Design, high-performance cloud computing as a data storage, mining, analysis of the core technology, the use of big data as a basis for organizational management decision support, make an overall plan and top-level design, construction and Education, management, service information as one of the key core database of school education information cloud service platform to form an information system suitable for the planning and development goals of the college and form an informationized learning environment and supporting service system for all to enjoy quality educational resources, Promote the teaching of online courses, e-resource sharing of book documents, virtual simulation experiment center and other supporting information construction, and promote the reform of mixed education teaching mode. We will promote the construction of information resources such as digital book resources and digital teaching resources, carry out extensive training in enhancing teachers 'informational instructional capabilities, and continuously improve teachers' information literacy. In the campus network infrastructure, data center infrastructure, information security and other aspects continue to develop, built into the office automation, educational administration, research management, student management, personnel management, enrollment and employment, logistics services, "Palm University" and other information subsystems One-stop online service hall to achieve the universal use of online resources in the teaching and learning process, to provide a solid information protection for the improvement of education and teaching quality, and to create a high-level intelligent campus model framework featuring information collection and resource sharing.

4. Conclusion

Although ant colony algorithm has obtained many improvements, there are many researches and applications in data clustering, but there are still some deficiencies that need further study: (1) How to
convert the actual clustering task into the problem of ant colony solving Space and express it in a right way. How to define the choice of "artificial ant" and the indirect communication between ants (such as the pheromone on the path, the distribution of objects, etc.); (2) The improper selection of parameters in the ant colony algorithm will directly produce the performance of the algorithm. Influence, but the method and principle of its selection are not yet theoretically basis, and can only be tuned through multiple experiments. Therefore, the parameter setting remains to be further studied; (3) The search time of ant colony algorithm is longer, only improvement and Improve the performance of the algorithm to adapt to large-scale data clustering; (4) ant colony algorithm mostly stays at the stage of experimental simulation. For its theoretical system, it still lacks a strict mathematical explanation, and its convergence proves to be incomplete. it will continue to develop.

References:
[1] Jiawei Han (Author)、 Micheline Kamber（Author）.Data Mining: Concepts and Techniques, Morgan Kaufmann（Publisher）, 2000.08
[2] P Berkhin.Survey of clustering data mining techniques[R].Technical report,Accure Software,San Jose,CA,2002
[3] A Dorigo,M Dorigo,V Maniezzo,Distributed optimization by ant colonies[C].In:Eurorpean Conference on Artificial Life,1991:134-142
[4] M Dorigo et al.Ant system:optimization by a colony of cooperating agents[j].IEEE Transactions on Systems,Man,and Cybemtics,Part B,1996;26(1):29-41
[5] E Bonabeau,M Dorigo,G Theraulaz.Swarm Intelligence-From Natural to Artificial System[M].New York,NY:Oxford University Press,1999
[6] E Lumer,B Faieta.Diversity and adaptation in populations of clustering ants[C].In:Proceedings of the Third International Conference on Simulation of Adaptive Behavior:From Animals to Animals,MIT Press,Cambridge,MA,1994:501-508
[7] J L Deneubourg,S Goss,N franks et al.Ths Dynamics of Collective Sorting:Robot-Like Ants and Ant-Like Robots[C].In:From Animals to Animates:Proceedings of the First International Conference on Simulation of Adaptive Behavior,MIT Press,1991:356-363
[8] M Parag Kanade,O Lawrence Hall.Fuzzy Ants as a Clustering Concept.Dept of Computer Science Engineering[C].In:22nd international conference of the North American fuzzy information processing society,NAFIPS,227-232