Research on Ant Colony Clustering Algorithm Based on HADOOP Platform

Zhihao Wang[1(✉)], Yonghua Huo[1], Junfang Wang[1], Kang Zhao[2], and Yang Yang[2]

[1] Science and Technology on Information Transmission and Dissemination in Communication Networks Laboratory, China Electronics Technology Group Corporation 54th Research Institute, Shijiazhuang, China
{cetc54001 6, jfwang2015}@sina.com, tsdhyh2005@163.com

[2] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China
489273711@qq.com, echo_lzjf@163.com

Abstract. Due to in the early period of the ant colony clustering algorithm convergence speed is very slow, this paper proposes a hybrid clustering algorithm based on ant colony clustering and MMK-means algorithm, which uses MMK-means algorithm to process the data, followed by ant colony clustering to finish clustering. Apart from that, this paper improves the ant colony clustering algorithm that makes ants using the best matching position, data object placement selecting and so on. We realize the algorithm in Hadoop platform, which can effectively reduce the time costs of clustering.

Keywords: Cluster · Hybrid algorithm · K-means · Ant colony algorithm

1 Introduction

Cluster analysis is a popular research branch of data mining. Clustering is the process that sets up the physical or abstract objects to similar objects composed of multiple classes or clusters. Clustering algorithms are applicable to the centralized data for clustering, while the actual data is distributed across different sites [1]. Due to the limit of transmission speed and safety factors, it is difficult to centralize the data in all sites to a central site. Besides, a large amount of data can result in a significant decrease on clustering efficiency.

Hadoop from the Apache open source community is an open source infrastructure software platform based on distributed system infrastructure which makes large-scale distributed computing and parallel processing widely used in cloud computing [2]. We propose a hybrid algorithm based on ant colony clustering and MMK-means

This work was supported by Open Subject Funds of Science and Technology on Information Transmission and Dissemination in Communication Networks Laboratory (ITD-U15002/KX15 2600011). NSFC(61401033,61372108,61272515). National Science and Technology Pillar Program Project (2015BAI11B01).

S. Wang and A. Zhou (Eds.): CollaborateCom 2016, LNICST 201, pp. 514–520, 2017.
DOI: 10.1007/978-3-319-59288-6_49

algorithm. This hybrid algorithm overcomes the shortcoming that ant colony clustering algorithm has slow convergence speed. In addition, we improve the ant colony clustering mechanisms that ants using the best matching position, data object placement selecting and so on.

The rest of this paper is organized as follows: In the second section, the related works are introduced; the method proposed in this paper is introduced in the third section; the fourth section is the analysis and comparison of the performance of several methods. The fifth section is the summary.

2 Related Work

K-means algorithm has the advantages that it has small computational efforts and fast convergence speed and it is suitable for large data processing [3, 4]. But this approach has obvious shortcomings. We need to determine in advance the number of initial cluster centers and clusters. It is greatly influenced by subjective factors. Besides, clustering results have poor stability and are prone to local optima.

MMK-means algorithm makes use of data sampling technique to take samples from the mass of data (This way can reduce the computational cost). MMK-means algorithm uses the largest and minimum distance algorithm to calculate the number of clusters that K-means algorithm requires and the initial cluster centers [5]. Then MMK-means algorithm makes use of K-means algorithm to cluster. It is clear that the result of MMK-means is better than K-means.

Ant colony algorithm is a novel bionic algorithm based on artificial swarm intelligence, which is inspired by the collective behavior of social insects. For real ants, they have two important types of behavior: foraging and clustering. The algorithms of ant colony optimization (ACO) have their origins in the ant foraging behavior [6–8]. They were proposed by Marco Dorigo and are useful in solving discrete optimization problems.

3 The Parallel Design of the Hybrid Clustering Algorithm

Each grid places a data object and each ant moves randomly in the grid. The ant calculates the neighborhood similarity $f(i)$, then decides the probability of picking up or dropping data object [8], the lower the probability to pick up when degree of similarity is higher, the greater the probability to drop.

$$f(i) = \begin{cases} \frac{1}{\delta^2} \sum_{j \in L} \left(1 - \frac{\delta(i,j)}{\alpha}\right) & if\ f(i) > 0 \\ 0 & otherwise \end{cases} \tag{1}$$

The neighborhood similarity $f(i)$ represents the average similarity of data i that the ants to pick up or drop with all the data objects in the visual radius r. $\delta(i,j)$ represents the distance between the data objects i and j. $\alpha \in [0,1]$ is the similarity adjustment coefficient; δ^2 is the size of neighborhood area L. Ant is in the center of the neighborhood, and neighborhood visual radius can be estimated according to (δ–1)/2.

If ants carry a data object at present, algorithm calculates the probability of ant put down the data P_d.

$$P_d = \begin{cases} 2f(i) & f(i) < k_d \\ 1 & f(i) \geq k_d \end{cases} \tag{2}$$

The ant compares P_d with the random probability P_r, if $P_d > P_r$,the ant puts down current carrying data and mark status to "un-carry" at the same time. Then the ant randomly selects a free data object, or the ant move to other random location. If ants do not carry a data object at present, algorithm calculates the probability of ant pick up the data P_p according to the following formula.

$$P_p(i) = \left(\frac{k_p}{k_p + f(i)}\right)^2 \tag{3}$$

The ant compares P_p with the random probability P_r, if $P_p > P_r$, the ant picks up data in the current location and mark status to "carry" at the same time. Then the ant moves to other random location, or let the ant randomly select a free data object.

Algorithm iterates until achieve the maximum number of iterations.

Here, we change the way that the ants always jump to the best matching position to the way that ant have different approaches in different stages. Improved algorithm commands ant move towards the direction of best matching position at the early stage. At the late stage ant makes a direct jump to the best match position. The ant needs to calculate the drop probability P_d when it moves to the new position. If $P_d > P_{r,}$ the ant drops carrying data with placement policy. Or else, it means that short-term memory is not valid, the ant moves to another location randomly. The ant should update its short-term memory when it success drops data.

When the ant put down a data object, if the ant find a suitable location to place s, but s was already occupied by other data objects, the ant colony clustering algorithm's solution is that the ant moves randomly to another location, but this solution has great blindness, slow convergence speed, and cannot effectively use the available information. When the ant wants to put down an object, this suggests that the data ant carried is similar enough with other data objects within a visual radius.

Improved steps for data object strategies are as follows: The ant searches for empty positions whose distance from the current position is 1, if there is an empty position then the ant drop the data object carried in this empty position. Otherwise ant searches for empty positions whose distance from the current position is 2, followed by recursion, until the distance reaches the visible radius. If there's still no empty position when the search radius reaches the visible radius, the ant moves to another position randomly, this can prevent the ant cannot put data object down in cases when the ant cannot find a proper empty position in the long iterative process. We make the rule that when the times that an ant fails putting down a data object reach the threshold we set, the ant finds the best matching position and drops a data object in the best matching position or best matching position's surroundings.

Improved ant colony clustering algorithm based on MapReduce framework [7] is shown in Fig. 1.

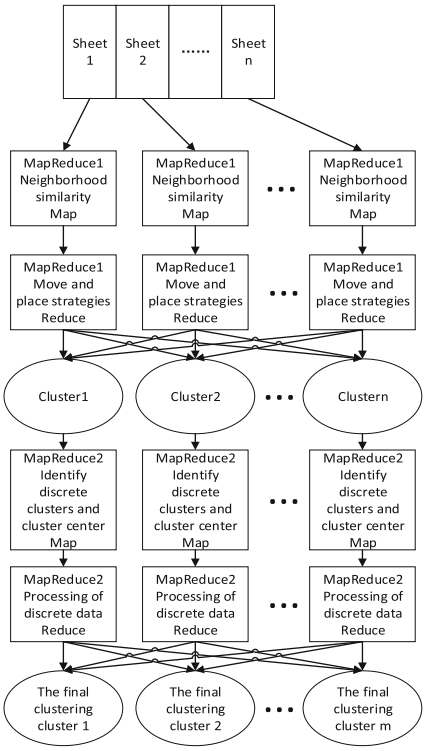

Fig. 1. Improved ant colony clustering based on MapReduce framework

(1) The original data set is divided into several pieces and distributed to each node. Map function of the MapReduce1 tasks is assigned to each node. Map function is responsible for mapping the data points in this node onto a two-dimensional grid randomly. Each Reduce function can be viewed as an ant, and responsible for the calculation of neighborhood similarity with Canberra distance. According to the number of iterations Reduce function make ants have variable movement speed (decreases in a random way), and along with the iteration ants' visual radius is increased gradually. If the case consistent with the "pick up principles" ant picks up data objects.

(2) When reaching the maximum number of iterations the algorithm enters MapReduce2 to deal with the clustering results. The clustering cluster is assigned to each node, the Map function in the MapReduce2 is responsible for comparing the number of data objects with threshold, if the number of data objects is greater than or equal to threshold, the cluster is identified as the correct cluster, if the number of data objects is less than the threshold, the cluster is identified as discrete cluster.

The algorithm calculates the center of all correct clustering clusters. Reduce function is responsible for calculating the distance of discrete data objects and the clustering center. If the distance conforms to the requirements of the threshold, the ant merge the data into the best match cluster, or mark the data as discrete data object.

(3) At last, algorithm output the final clustering results.

The hybrid clustering algorithm is shown in the Fig. 2. We first use of MMK-means algorithm's fast convergence characteristics to preprocess data. This section includes data sampling, using the maximum minimum distance method to select the cluster centers in the extracted data and merging adjacent center to obtain the initial cluster centers for K-means algorithm. After K- means algorithm clusters roughly by using the initial clustering center and clustering cluster number. At last, the algorithm uses improved ant colony clustering algorithm to cluster.

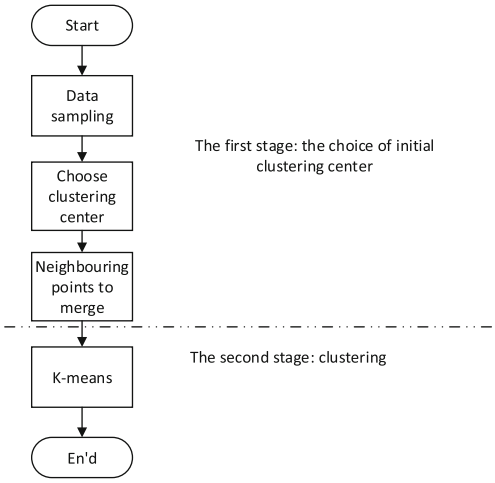

Fig. 2. The hybrid clustering based on MMK-means and ant colony clustering

4 Experiment and Analysis

Simulation data set's each property is in accordance with normal distribution. Dimension property is two-dimensional. Data set consists of five classes, and data for each class are in line with the Gaussian distribution. There are 1000 data objects in total. The simulation experiments in this article use the following parameter settings, and we take 30 experiments with the data set:

$P_p = 0.1$, $P_d = 0.1$, $\alpha = 0.1$, N = 100, short-term memory capacity of ant $N_{memory} = 20$, the number of ant $N_{ant} = 10$.

Figure 3 shows the clustering result after using improved ant colony clustering algorithm. The number of discrete points is significantly less than that generated by

traditional ant colony clustering algorithm, due to the introduction of the discrete data processing strategy. Figure 4 shows the clustering result using the hybrid clustering algorithm based on MMK-means algorithm and ant colony clustering algorithm. The clustering effect is similar with that generated by improved ant colony clustering algorithm, because they use the same principle of clustering.

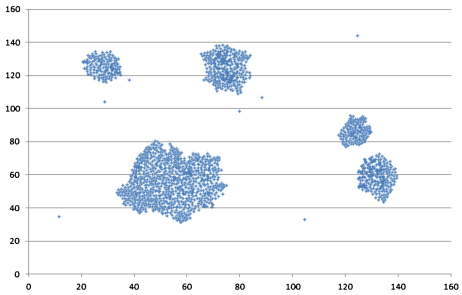

Fig. 3. The clustering result after using improved ant colony clustering algorithm

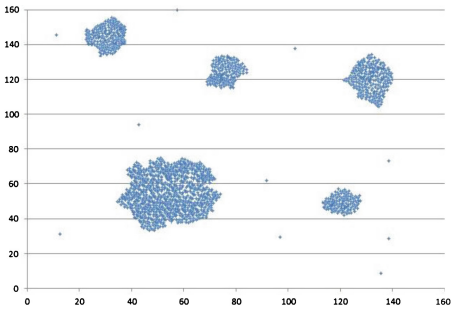

Fig. 4. The clustering result after using the hybrid clustering algorithm

F-measure is also known as F-score, the most common evaluation criteria in the field of information retrieval [1, 2]. Figure 5 shows the F-measure value of the five algorithms described in this article. The F-measure value of the hybrid clustering is superior to that of the improved ant colony clustering. The high-to-low order of accuracy of the clustering effect reflected by F-measure values is ant colony clustering, MMK-means and K-means. Improved clustering is better than ant colony clustering, mainly because the improved algorithm has adaptive characteristics.

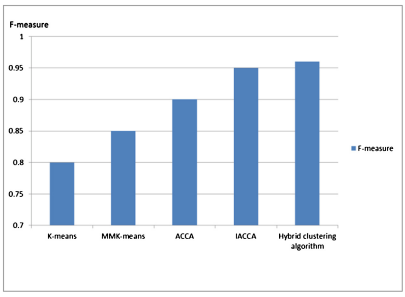

Fig. 5. F-measure of various clustering algorithms

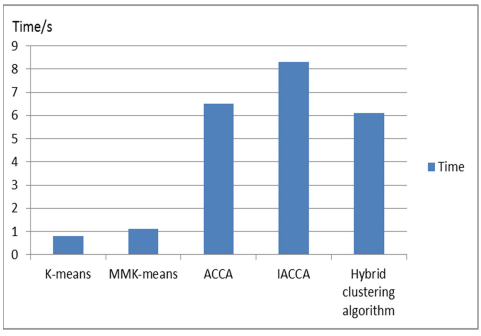

Fig. 6. Time performance comparison of various clustering algorithms

Figure 6 shows the time performance comparison of various clustering algorithms. The time performance of the hybrid clustering is better than that of the improved ant colony clustering algorithm (IACCA) with decreasing by 26.2%. By sacrificing some time cost, it improves the effectiveness of clustering. MMK-means decreases the time performance by 50.8% than K-means, which is due to that MMK-means needs to calculate the additional parameters that the K-means algorithm needs. The hybrid clustering algorithm improves the time performance by 5.7% than the basic ant colony clustering algorithm (ACCA) and improves the time performance by 25.6% than IACCA. So the hybrid clustering algorithm can improve the convergence speed.

5 Conclusion

The proposed parallel design of MMK-means algorithm runs in parallel on Hadoop platform. Improved ant colony clustering algorithm and its parallel design effectively improve the clustering effect and the time cost. Finally, the mixed clustering algorithm accelerates the convergence speed and improves the clustering effect compared with ant colony clustering algorithm.

References

1. Wei, X.: Clustering algorithm based on the combination of genetic algorithm and ant colony algorithm. In: International Conference on Innovative Computing & Cloud Computing, pp. 45–49. ACM (2011)
2. Kenidra, B., Meshoul, S.: A data-clustering approach based on artificial ant colonies with control of emergence. In: Soft Computing and Pattern Recognition, pp. 430–435. IEEE (2014)
3. Asbern, A., Asha, P.: Performance evaluation of association mining in Hadoop single node cluster with big data. In: International Conference on Circuit, Power and Computing Technologies. IEEE (2015)
4. Jiang, H., Zhang, G., Cai, J.: An improved ant colony clustering algorithm based on lf algorithm. In: 2015 IEEE 12th International Conference on e-Business Engineering (ICEBE), pp. 194–197. IEEE Computer Society (2015)
5. Yu, H., Wang, D.: Mass log data processing and mining based on Hadoop and cloud computing. In: International Conference on Computer Science & Education, pp. 197–202 (2012)
6. Zhou, A., Wang, S., Sun, Q., et al.: Dynamic virtual resource renting method for maximizing the profits of a cloud service provider in a dynamic pricing model. In: International Conference on Parallel and Distributed Systems, pp. 944–945 (2013)
7. Wang, S., Zhou, A., Hsu, C.H., et al.: Provision of data-intensive services through energy- and QoS-aware virtual machine placement in national cloud data centers. IEEE Trans. Emerg. Top. Comput. **4**(2), 1 (2015)
8. Mao, L., Shen, M.M.: An improved ant colony clustering algorithm based on dynamic neighborhood. In: IEEE International Conference on Intelligent Computing and Intelligent Systems, pp. 730–734 (2010)