Research on the Algorithm of Data Mining Based on Cluster Analysis
Algorithm of K-means Based on the Pheromone

Zhang Qun
School of Computer Science
Hubei University of Technology
Wuhan, China
184034298@qq.com

Wang Chun-zhi
School of Computer Science
Hubei University of Technology
Wuhan, China
mylostapple@sohu.com

Abstract—Cluster analysis becoming increasingly essential in data mining field, and is mainly used to discover the valuable data distribution and data mode in the potential datum. Based on the pheromone studies on basic clustering model, the theory of information entropy and two classical clustering analysis algorithms, an algorithm of K-means based on the pheromone is presented firstly. The algorithm works with the transformation probability to realize the clustering. It has reduced the number of the parameters and improved the speed of clustering.

Keywords—ant colony; data mining; K-means algorithm

I. INTRODUCTION
Since the presence of ant colony algorithm (ACA) provided by M.Dorigo et al in 1991, many researchers are attracted to make the research on it, and apply it to solve the combinatorial optimization issues successfully, such as TSP(Traveling Salesman Problem), QAP(Quadratic Assignment Problem), JSP(Job-shop Scheduling Problem), etc. ACA is used to solve many combinatorial optimization problems. As long as the diagram can be made out to explain the problems to be solved, then positive feedback process (such as the residual information in TSP) can be defined, and the problem structure can provide heuristic information (such as inter-city travel fee in TSP) by itself and the establishment of constraint mechanism (such the visited city list in TSP). Several years ago, Ant Colony Optimization (ACO) algorithm includes the ACA provided by M.Dorigo et al in 1991, and from then on, many related algorithm frames has been provided.

Some foreign scholars have made contributions on the ACA since last decade, where K-means algorithm is one of the clustering algorithms that are used the most widely.

II. K-MEANS ALGORITHM
K-means algorithm was presented by MacQueen firstly, which belongs to the means based on the pheromone partition in clustering means. It is the clustering algorithm given the clustering category, and is the simplest iterative optimization method.

Clustering algorithms based on the pheromone partition are mostly described as: give d --- dimensional space $R^d$; in $R^d$ define an evaluation function $C : \{X : X \subseteq S\} \rightarrow R^+$ to make quantified evaluation for each cluster; input object collection $S$ and one integer $k$ in $R^d$; require to output a partition of $S$: $S_1$, $S_2$, ..., $S_k$; The partition minimizes the objective value function $J$. In which:

$$J = \sum_{i=1}^{k} C(S_i)$$  \hspace{1cm} Formula 1

Different evaluation functions will result in different clustering results. Mostly, the sum of squared Euclidean distance between any two objects in all clusters is considered as clustering evaluation function. The function is defined as below:

$$C(S_i) = \sum_{i=1}^{n} \sum_{s=1}^{n} d^2(x_i, x_s)$$  \hspace{1cm} Formula 2

In which, $S_i$ means the cluster resulted from partition; $|S_i|$ means the number of the objects, $x_i$, $x_s$ means the rth and sth objects in $S_i$ respectively; $d(x_i, x_s)$ means the Euclidean distance between the two objects of $x_i$ and $x_s$.

But in K-means algorithm, the sum of squared Euclidean distance between all sample objects and their centers is considered as clustering evaluation function. Each cluster $S_i$ center is also the mean value of the objects in cluster $S_i$. After each clustering process, it will be calculated again and enter the next round of clustering as the new cluster center, until the cluster center won’t change again.

$$C(S_i) = \sum_{i=1}^{n} d^2(\bar{x}_i, x_i)$$  \hspace{1cm} Formula 3

In which, $\bar{x}_i$ is the center of cluster $S_i$. The meaning of other symbols is just like Formula 2.

K-means algorithm is defined as below:

(1) Randomly select $k$ objects to be the original cluster center.
(2) Repeat the steps of (3), (4), until the cluster center won’t change again.
(3) Calculate the distance between all objects and the cluster center, and sort the objects to the cluster which is the nearest to the objects.
(4) Calculate the new cluster center for the adjusted new cluster. It means to re-calculate the mean value of the objects in all clusters, and make the object mean value point as new cluster center.

K-means algorithm is the classical algorithm to solve clustering analysis problems. It not only generates $k$ clusters,
but also calculates the center of each cluster. K-means algorithm has good stretch and high efficiency, but traditional K-means algorithm has two inherent defects:

1. The selection of original random cluster center would result in different clustering results, or even no solution.
2. The algorithm is hiking-type search algorithm based on the pheromone gradient descent, so it is inevitable to be lost in partial extreme value.

Both defects dramatically limit its application. For the first defect, the improvement solution will be explained carefully in part four in this paper; for the second defect, this paper comes up with the improved K-means algorithm based on pheromone.

### III. K-Means Algorithm Based on Pheromone

In the process of ant foraging, pheromone determines their move orientation. Therefore, this paper comes up with the improved K-means algorithm based on the pheromone. The algorithm considers data objects as the ants of different natures, and considers the cluster center as the “food source” for ants, so the clustering process can be considered as the process of ant foraging.

K-means algorithm based on the pheromone employs the feature of ant clusters’ distributed search (don’t converge too early), so the defect that traditional K-means algorithm is easy to loss in extreme value will be improved. The algorithm concept is: introduce the ant transition probability from food source i to food source j into K-means algorithm. The attribution of data objects is determined by the transition probability. Introduce standard of cluster deviation into the next round of circulation, update the cluster center, calculate the deviation, and make the judgment again, until the deviation has no change or is within certain error range. Then the algorithm is finished.

The algorithm is as below:

If: \( X = \{X_i | X_i = (x_{i1}, x_{i2}, \ldots, x_{in}), i = 1,2,\ldots,n \} \) is the data set to be clustered;

If: \( C_j \) is the cluster center, original value is distributed evaluation;

At the time \( t \), the pheromone quantity \( \tau_{ij}(t) \) on the ant route \((i, j)\) from \( i \) to the food source \( j \) (cluster center) is defined as:

\[
\tau_{ij}(t) = \begin{cases} 
1, & d_{ij} < R \\
0, & d_{ij} \geq R 
\end{cases}
\]

In which, \( d_{ij} \) means the Euclidian distance between object \( X_i \) and the cluster center \( C_j \), \( R \) is pre-set cluster radius; If \( \tau_{ij}(0) = 0 \), the transition probability object \( X_i \) to the cluster center \( C_j \) is determined by \( P_{ij} \):

\[
P_{ij}(t) = \frac{\tau_{ij}(t) \cdot \eta_{ij}(t)}{\sum_{k=1}^{n} \tau_{ik}(t) \cdot \eta_{ik}(t)}
\]

In which, \( S = \{X_i | d_{ij} \leq R, s = 1,2,\ldots,n \} \) means the set of data objects sorted to the cluster center \( C_j \). If \( P_{ij}(t) \geq P_o \) (\( P_o \) is given), \( X_i \) sorts to \( C_j \); otherwise, it would not.

If \( CS_j = \{X_i | d_{ij} \leq R, i = 1,2,\ldots,J \} \) means the set of data objects sorted to the cluster center \( C_j \), and \( J \) is the number of the objects in the set, then the cluster center can be calculated by Formula 6. The deviation error \( \xi_j \) of the ith cluster and the total deviation error \( \bar{\xi} \) of all clusters can be calculated by Formula 7 and Formula 8.

\[
C_j = \frac{1}{J} \sum_{i=1}^{J} X_i, \quad \text{Formula 6}
\]

\[
\xi_j = \frac{1}{J} \sum_{i=1}^{J} (X_i - C_j), \quad \text{Formula 7}
\]

\[
\bar{\xi} = \sum_{j=1}^{J} \xi_j, \quad \text{Formula 8}
\]

The process diagram of the K-means algorithm based on pheromone is just like the following Figure 1.

**Figure 1.** The process diagram of K-means algorithm based on the pheromone information entropy

### IV. Comparison Analysis on Both Algorithms

K-means algorithm clusters according to the distance between objects, while the K-means algorithm based on the
pheromone clusters on the transition probability. It is the biggest difference between both algorithms. The clustering performance of both algorithms will be analyzed through the example of 2–dimensional data set of ten random generated numbers within (0, 1), just as below:

If the data set for clustering \( X = \{(0.44,0.28), (0.05,0.42), (0.77,0.31), (0.41,0.99), (0.72,0.83), (0.97,0.42), (0.18,0.80), (0.42,0.02), (0.14,0.62), (0.98,0.03)\} \), and define them in order as \( X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8,X_9,X_{10} \);

If \( k = 3; \ R = 0.35; \ P_0 = 0.4; \ \epsilon_0 = 0.6. \)

The randomly set cluster centers are \( X_2, X, X_8. \) Then find the optimal clustering result by K-means algorithm (See Figure 2 for the clustering result) and the K-means algorithm based on the pheromone (See Figure 3 for the clustering result).

![Figure 2. Clustering result of the ten sets random samples by K-means algorithm](image)

![Figure 3. Clustering result of the ten sets random samples by the K-means algorithm based on the pheromone](image)

According to the above figures (the same shape means to be marked as the same cluster) and Table 1 (Table 1 lists specific clustering results and the deviation value of all clusters), obviously the deviation value of the K-means algorithm based on the pheromone is less than the final value of K-means algorithm. To some extent, it confirms that the algorithm is reasonable.

<table>
<thead>
<tr>
<th>N</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>( \epsilon )</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0.626632</td>
</tr>
<tr>
<td>K-means Based on Pheromone</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0.579590</td>
</tr>
</tbody>
</table>

V. CLUSTERING COMBINATION ALGORITHM BASED ON ACA

A. Basic concept

Clustering combination algorithm based on the pheromone ACA can be considered as the combination of ACA based on information entropy and the K-means based on the pheromone. ACA based on information entropy originates from the basic ant colony clustering model and Luhnner and Faieta’s data analysis LF algorithm. ACA based on information entropy is the self-organizing clustering algorithm with the features of robustness and visualization. It can generate some new and meaningful clustering models. The K-means algorithm based on the pheromone is the improved algorithm of K-means. The K value and the original clustering center in the K-means algorithm are given previously, but both selections will affect the clustering effect and speed directly (one the defects in K-means algorithm), therefore, according to and with the combination of the K-means algorithm based on information entropy, clustering combination algorithm is provided. The algorithm can be carried out by three steps:

Phase 1: Clustering process of ant colony clustering algorithm based on information entropy.

A colony of ants randomly moves in the 2-dimensional area, and they can pick up and put down any objects. Before or after the calculating data was picked up or put down, the information entropy judges whether the data objects should be picked up or put down through the comparison of both algorithms. Through the training repetition, this step will generate more proper clusters.

Phase 2: Collection and identification of data samples

This step refers to the collection of single ACA (SACA) and the method to express data samples. The method searches and recursively collects all data related to samples.

Phase 3: Clustering optimization process for the K-means algorithm based on the pheromone

Take the above cluster center and the number of clusters as the original value, and consider the data set as ant colony, the cluster center as food source, and then cluster for the second time.

Abbreviations in the title or heads unless they are unavoidable.

B. Algorithm description

1) Main process diagram (Figure 4) of the algorithm.
2) Algorithm process description.

a) Step 1: Parameter initialization. Set the maximum number of cycles max_cn, the number of ants Nant, the number of minimum data objects in the cluster min_no, /g68/, Nant, \beta, cluster radius R, P0, \epsilon_0, etc.

b) Step 2: Randomly distribute the data objects for clustering to the 2-dimensional area. Give each data object one coordinate (Xi,Yi) randomly, in which, Xi,Yi \in [S,S], i = 1,2,...,n.

There are two important data structures:
- struct DataType // Data object structure
  {  double *data;
      int father;
      double *uncle;
  };

- struct ClusterType // Cluster structure
  {  double *center;
      int son_num;
  };

c) Step 3: Randomly put ants into the 2-dimensional area, and set the original condition for them. Initial state has no load.

d) Step 4: Ants move randomly, repeat the following steps to all ants until the repetitions reach max_cn:

(1) If this ant has no load and reaches the area of one data object, then use LF improved algorithm based on the information entropy to calculate the data object information entropy E1 in the area of SxS when the ant doesn’t pick up the data object and to calculate the information entropy E2 when the ant picks up the data object. If E1>E2, then the ant puts down the object, and it will have no load and new coordinate; otherwise, the ant would not put down the object.

(2) If the ant has load and reaches the empty area, then use LF improved algorithm based on the pheromone to calculate the data object information entropy E1 in the area of SxS when the ant doesn’t put down the data object and to calculate the information entropy E2 when the ant puts down the object. If E1>E2, then the ant puts down the object, and it will have no load and new coordinate; otherwise, the ant would not put down the object.

e) Step 5: Refer to the method of SACA mark data samples to mark the data. If the data object is not marked, then carry out the marking operation. If the data object is alone or the number of data objects with its field is less than min_no, then mark the data object as the lonely point; otherwise, the data object would be offered a cluster serial number, which is the same with the number of other data objects in other field.

f) Step 6: Cluster optimization of K-means algorithm based on the pheromone:

(1) Calculate the mean value of all clusters centers, except the lonely points;
(2) Use Formula 5 to calculate the transition probability P_{ij} between of each lonely data object i and the cluster center C_j. If P_{ij}>=P_0, then the lonely data object will be sorted to the belonging cluster.

(3) Calculate the deviation error \xi_j of each cluster and the total deviation \xi of all clusters, and update the cluster center and information amount \tau_{ij}.

(4) The program will end, until the departure exceed \xi \leq \xi_0.

g) Summary: Step 1 to 3 is the initialization phase for the whole combination algorithm. Its main functions are to initialize the process and to distribute data objects randomly in the 2-dimensional area; Step 4 realizes the clustering process by ACA based on information entropy. Its function is to form better clusters; Step 5 is the marking process for data objects, namely the collection process of clustering results; Step 5 improves clusters by the K-means algorithm based on the pheromone. The previous results are the preconditions of this process.

VI. PERFORMANCE ANALYSIS ON THE COMBINATIONAL ALGORITHM

When the paper researches classical K-means algorithm, it comes up with the K-means algorithm based on the pheromone, which is the improved K-means algorithm combining the search capability in ant foraging. This paper test the algorithm through a pair of artificial data, and the result is that under the same conditions, the deviation value of the K-means algorithm based on the pheromone is less than K-means algorithm. It means that improved algorithm has better clustering results. However, the K-means algorithm based on the pheromone has uncertain elements. The uncertainty in k value (including k value and original
cluster center) will result in the uncertainty of final clustering results.

For clustering combination algorithm, the paper provides the solution to use the LF method based on information entropy to cluster (the first cluster), and use the K-means algorithm based on the pheromone to improve the cluster results (the second cluster). The solution uses the initial value of the first cluster and the second cluster, so the uncertainty caused by the uncertainty in k value will be avoided. The combination algorithm has better performance than single algorithm.

VII. CONCLUSION

To solve the defect of traditional K-means algorithm, this paper comes up with the improved K-means algorithm based on the pheromone by the inspiration of ant foraging principle. Associate the data cluster from ant foraging, consider the data objects as the ants, the cluster center as the food source, and use the transition probability to sort data objects. Experimental results show that the algorithm is proper and effective. Then the paper provides the SACA according to the K-means algorithm based on the pheromone. The solution firstly uses the LF method based on information entropy to cluster, and then uses the K-means algorithm based on the pheromone to improve the cluster results.

But it still has some defects:
1. The update system of the information entropy in the K-means algorithm based on the pheromone doesn’t give the consideration on the volatilization with the time. It is the content needed to be improved in the next research.
2. The solution of the clustering combination algorithm on the basis of ACA is applicable in theory, and the establishment of such system is available in technology, but the application research and specific practice are not carried out because of limited knowledge and data values.
3. ACA is widely used to research the route. As the time concerns, it doesn’t make the deep research on the ant colony route algorithm. So the research on the ant colony route algorithm will be carried out in future.

ACKNOWLEDGMENT

Thanks for my parents take good care of me, and thanks for my friends care for me.

REFERENCES