User Clustering Balance and Quantization Scheme Based on Data Mining

Jinzhu Li¹, Yuan Yuan¹, Yujie Zhang⁺,¹,²

¹ School of Information Science and Engineering, Dalian Polytechnic University, Dalian, 116034, Liaoning Province, China.
² School of Food and Technology, Dalian Polytechnic University, Dalian, 116034, Liaoning Province, China.

Yujie Zhang⁺, Corresponding Author, Associate Professor, Email: zhangyj@dlpu.edu.cn

Abstract. The problem studied in this paper is how to balance the merits and demerits of all tasks in labor crowdsourcing and quantify the overall completion degree under various factors, and promote the solution to maximize the task completion with joint packaging. This paper introduced a data mining model to re-plan all task distribution scatters to solve such problems. Firstly, process a data cleaning, screen out valid data and analyze factors in natural situations according to the K-means clustering algorithm. Then, preprocess data by discrete point detection, weigh the pros and cons by the FCM clustering algorithm, and calculate the maximum task completion degree. Finally, by both, optimize initial parameters by iterative calculation while packaging with DBSCAN clustering algorithm, after that, substituting the optimal result cluster under the mean constraint to obtain a common packing scheme. This paper is based on a variety of data mining models to reach a prominent solution and a closely combined algorithm which can accurately analyzing the advantages and disadvantages, and can even calculating the degree of task completion to test the solution. The result has distinct advantages and easy to use.

1. Introductions

Nowadays, in the information age, most people own their mobile devices, so many new services come into being. According to people's dispersed location, functions of equipment and the information they have access to, the network is used as the carrier to make the effective information flow on the Internet. Therefore, the paid estimation and timely distribution processing of these effective information is the core of the service [1]. The models mentioned in this paper can simulate the distribution of tasks reliably, and also can efficiently carry out thorough analysis and calculation.

We first analyzed and evaluated the natural state data, then established the optimization model, finally proposed the solution with the maximum overall completion degree, and set up the following basic model framework:
Before we start modeling, in order to screen out valid data, we need to conduct data cleaning first, aiming at removing outliers, correcting existing errors and providing data consistency [2].

K-Means clustering algorithm is a typical distance-based non-hierarchical clustering algorithm, on the basis of minimizing the error function, the data is divided into a predetermined class number K, and the distance is used as the evaluation index of similarity, that is, the closer the distance between two objects is, the greater the similarity will be [3]. Based on this algorithm, the feature analysis of naturalized sequences can be carried out.

In order to propose an optimization method, outlier detection is an effective method to select outliers based on K-means, which is helpful to enumerate the worst cases and the focus can be put on the study of a more cohesive point set [2].
FCM mean clustering algorithm is a fuzzy clustering algorithm based on objective function [4]. With its fuzzy characteristics, we re-cluster the condensed points to improve the degree of completion and form a new optimization method.

Finally, a higher-level model is considered based on the two methods. The DBSCAN algorithm is the most widely used density-based clustering analysis algorithm [5]. Different from the partitioning and hierarchical clustering methods, cluster is defined as the maximum set of points connected by density in this model, so that it can divide the area with sufficient density into clusters and find clusters of arbitrary shapes including noise points in the spatial database of noise. From the original scattered and disorderly points to the compact point clusters with related cohesion, the solution to joint packaging and distribution of tasks can be expanded and promoted to maximize the degree of completion.

2. Assumptions
The number of members does not change in a short time.
All task points have only two states, completed and unfinished.
Tasks are automatically assigned by the service system, and there are no duplicate completion and competition conditions.

3. Symbols

| No. | Symbol      | Description                                                                                                                  |
|-----|-------------|-----------------------------------------------------------------------------------------------------------------------------|
| 1   | GPS − Xi    | Latitude of the task in the project                                                                                        |
| 2   | GPX − Yi    | Longitude of the task in the project                                                                                       |
| 3   | S           | The status of the task completion (1 is completed, 0 is not completed)                                                        |
| 4   | R           | Credit value of the member                                                                                                  |
| 5   | price       | Price of the task                                                                                                           |
| 6   | M           | Scheduled task limit of the member                                                                                         |
| 7   | GPS − Xm    | Latitude of the member                                                                                                       |
| 8   | GPX − Ym    | Longitude of the member                                                                                                      |
| 9   | L(MaxC − Center) | The largest Euclidean distance from the point in the cluster center to the cluster center                                  |
| 10  | L(MinC − Center) | The smallest Euclidean distance from the point in the cluster center to the cluster center                                 |
| 11  | C(MaxC)     | The point near the farthest point from the center of the cluster that has the highest price completed                            |
| 12  | C(MinC)     | The point near the farthest point from the center of the cluster that has the lowest price completed                            |

Fig. 3 Discrete point detection
4. Models

4.1. K-means clustering model

First, given a data set $X$ containing $n$ data objects, the number of clusters is $k$, $c_1, c_2, c_3 \cdots c_k$ are initial clustering centers, $k$ categories are represented by $W_1, W_2, W_3 \cdots W_k$, the number of data objects of $W_i$ is represented by $n_i$ [3].

Enter the number $k$ of data sets to be classified (after data cleaning) and number of clusters $k$, then initialize the cluster centers.

Use the selected distance calculation method (this method takes the Euclidean distance), calculate the distance $d$ from each data object in the data set $D$ to the cluster center, if

$$d(x_i, c_j) < d(x_i, c_m), \quad i = 1, 2, \cdots, k, \quad j = 1, 2, \cdots, k, \quad m = 1, 2, \cdots, k, \quad j \neq m$$

Then $x_i$ is assigned to $c_j$, otherwise returned to class $c_m$.

Common distance calculation methods:

Minkowski distance:

$$d_q(q) = \left( \sum_{a=1}^{p} |x_{ia} - x_{ja}|^q \right)^{1/q}$$

When $q = 1$:

$$d_1 = \sum_{a=1}^{p} |x_{ia} - x_{ja}|$$

is Absolute distance

and $q = 2$:

$$d_2 = \left( \sum_{a=1}^{p} (x_{ia} - x_{ja})^2 \right)^{1/2}$$

is Euclidean distance

as $q = \infty$

$$d_\infty = \max_{1 \leq a \leq p} |x_{ia} - x_{ja}|$$

is Chebyshev distance

According to the formula:

$$c_i = \frac{1}{n} \sum_{x \in w_i} x, \quad (j = 1, 2, \cdots, k)$$

Recalculate the center of the clusters:

$c_1^*, c_2^*, c_3^* \cdots c_k^*$

If for any $i \in \{1, 2, \cdots, k\}$ satisfies $c_i^* = c_i$, then the algorithm ends, $c_1^*, c_2^*, c_3^* \cdots c_k^*$ are the final $k$ classes, or let $c_i = c_i^*$. The algorithm continues from b);

Output clustering results.

4.2. FCM mean clustering model

Divide $N$ $L$-dimensional vectors into $C$ fuzzy groups, through iteratively updating of membership and clustering center, clustering data by minimizing objective function last [6].

Where, the objective function is:

$$J(U, V) = \sum_{i=1}^{N} \sum_{c=1}^{N} u_{ic}^m d^2(x_i, v_c)$$

(1)
The constraint condition is:

$$\sum_{c=1}^{C} u_{ic} = 1 \quad \forall i$$

According to the nonnegativeness of membership, there is $$u_{ic} \geq 0 \quad \forall i,c$$ , and $$\sum_{i=1}^{N} u_{ic} > 0 \quad \forall c$$.

Here, $$m$$ refers to the fuzzy weighting coefficient, It's greater than 1; $$d(x_i, v_c)$$ represents the Euclidean distance between the first data point and the c-th cluster center; $$u_{ic} \in [0,1]$$ is an element in the membership matrix; $$v_c$$ is the clustering center corresponding to each cluster;

In order to find the extremum of the objective function with constraints, a Lagrangian factor is introduced to construct a new objective function:

$$J_\lambda(U,V) = \sum_{i=1}^{N} \sum_{c=1}^{C} u_{ic} m d^2(x_i,v_c) + \lambda (\sum_{c=1}^{C} u_{ic} -1)$$

The optimization conditions for the extremum of the objective function are as follows:

$$\frac{\partial J_\lambda}{\partial \lambda} = (\sum_{i=1}^{N} \sum_{c=1}^{C} u_{ic} -1) = 0$$

$$\frac{\partial J_\lambda}{\partial u_{ic}} = (\sum_{i=1}^{N} \sum_{c=1}^{C} u_{ic} m d^2(x_i,v_c) - \lambda) = 0$$

$$\frac{\partial J_\lambda}{\partial v_c} = (\sum_{i=1}^{N} \sum_{c=1}^{C} u_{ic} m x_i - v_c \sum_{i=1}^{N} u_{ic} m) = 0$$

Thus, the calculation formulas of membership degree $$u_{ic}$$ and clustering center $$v_c$$ can be obtained as follows:

$$u_{ic} = \frac{1}{\sum_{j=1}^{C} \left( \frac{d(x_i, v_c)}{d(x_i, v_j)} \right)}$$

$$v_c = \frac{\sum_{i=1}^{N} u_{ic} x_i}{\sum_{i=1}^{N} u_{ic} m}$$

According to the above formulas (1)-(7), the membership degree and clustering center satisfying the conditions can be figured out through continuous iteration [7].

The specific steps of the algorithm are as follows:
Set the number of categories C and the fuzzy coefficient $$m$$;
Initialize the membership matrix and satisfy the normalization condition of the formula (3);
Calculate the clustering center according to formula (6);
Update the membership matrix according to formula (5);
Compare the iterative membership matrix to the matrix norm. Stop iterate if $$\|U^{(t)} - U^{(t-1)}\| < \varepsilon$$ , otherwise return step c).
Because the value of the fuzzy matrix conforms to the normal distribution, we tested and analyzed the confidence interval of $$2\sigma$$ after obtaining the fuzzy sets. Finally, combined with the new pricing function, a new scheme.
There is the pricing function:
\[ L(\text{MaxC} - \text{Center}) = k(C(\text{MaxC}) - C(\text{MinC})) \]

4.3. **DBSCAN clustering model**

Input: Sample set \( D = (x_1, x_2, \cdots, x_m) \), Neighborhood parameter \((\varepsilon, \text{MinPts})\), Sample distance is calculated by using Euclidean distance measure;

- Initialize the core object collection \( \Omega = \emptyset \), clustering cluster number \( k = 0 \), unvisited sample set \( \Gamma = D \) and cluster partition \( C = \emptyset \);
- Toward \( j = 1, 2, \cdots, m \), find all core objects by following the steps below:
  - Find \( \varepsilon \)- neighborhood sub-sample sets \( N_\varepsilon(x_j) \) of sample \( x_j \) by using distance measure;
  - If the number of samples in the subsample set satisfies \( |N_\varepsilon(x_j)| \geq \text{MinPts} \), incorporate sample \( x_j \) into the core object sample collection: \( \Omega = \Omega \cup \{x_j\} \).
  - If the core object set satisfies \( \Omega = \emptyset \), the algorithm ends, otherwise it goes to the next step;
- Randomly select a core object \( O \) from the core objects, initialize the current cluster core object queue \( \Omega_{\text{cur}} = \{O\} \) and update unvisited sample collection \( \Gamma = \Gamma - \{O\} \).
- The current clustering cluster \( x \) is generated when the current cluster core object queue \( \Omega_{\text{cur}} = \emptyset \), renew cluster partitions \( C = \{C_1, C_2, \cdots, C_k\} \) and the core object set \( \Omega = \Omega - C_k \).
  - Here, execute step g) after make sure step d) in the previous step, or continue the follow step.
- Extract a core object \( O \) in the current cluster core object queue \( \Omega_{\text{cur}} \), and find out all \( \varepsilon \)-neighborhood subsample set \( N_\varepsilon(O) \) by neighborhood distance threshold \( \varepsilon \).
  - Where, \( \Delta = N_\varepsilon(O) \cap \Gamma \), return to step f) after updated the current cluster sample set \( C_k = C_k \cup \Delta \) and the unaccessed sample collection.
  - The output result is: cluster partitions \( C = \{C_1, C_2, \cdots, C_k\} \).

5. **Models validation**

The data used in this model test is the actual data provided by Paiba Technology (Tianjin) Co., Ltd.

5.1. **Natural situation analysis**

The clustering center obtained by K-means cluster analysis using Matlab is shown in Table 1.

| Number | Parameter | 1     | 2     | 3     |
|--------|-----------|-------|-------|-------|
|        | GPS-X     | 0.583718 | -0.5998 | -0.49394 |
|        | GPS-Y     | 0.346483 | -0.6205 | 1.571503 |
|        | price     | -0.9559  | 1.121663 | -0.16484 |

And the stereoscopic figure:
It can be seen that each type of price distribution has a large difference: most of the unfinished tasks are distributed between [65, 67.5] but the price of the completed task is evenly distributed at [65, 75], the clustering green cluster center is surrounded by unfinished tasks prove that the pricing of the task is a decisive factor. That is to say that many tasks are not completed mainly due to these are underpriced when members choosing them.

5.2. Balance and quantity optimization

Run MATLAB program, the calculation of FCM model is as follows [8]:

| Parameter | 1   | 2    |
|-----------|-----|------|
| GPS-X     | 0.5837 | -0.5998 |
| GPS-Y     | -0.5998 | -0.6231 |
| price     | -0.4939 | 1.5715 |

According to the characteristics of the algorithm feature value and the new pricing function, the maximum value of each column in the obtained fuzzy partition matrix is taken as a normal distribution of 2σ, of which 89 are rejected. Finally, we believe that the biggest task can be completed 731, better than the number 522 in a natural state.
5.3. **Completeness maximization scheme**

Machine learning training results are as follows:

62 36 32 26 21 18 12 6 5 5 5 19
19 19 5 5 5 5 5 5 5 4 4 4

If the initial value is too small, the noise is too high; if the initial value is too large, the error is very large. Thus, we should take the 10th value (the 9th number is 5) as the optimal solution.

Run the DBSCAN algorithm program to get 93 clusters of \{5 3 10 13 7 85×9 11 14 7\}, which is the number of tasks in each package.

Like is:

Fig. 6 DNSCAN result diagram

6. **Model evaluation**

This model does not directly perform regression, curve fitting, etc. Avoided the low degree of fitting makes the analysis inaccurate due to lots of unrealistic data. But a different approach, using the combined clustering algorithm based on Data Mining, which not only gives a way to weigh the pros and cons, but also quantifies the overall degree of completion to give a solution, and also expands to a program that uses joint packaging to achieve maximum completion.

7. **Improvement ideas**

Since the mere consideration of the nature of the task itself affects itself and ignores many external factors, the analysis of Problem 2 is somewhat idealized.

**References**

[1] Liu Xiaogang. Factors Affecting task publisher to bid in crowdsourcing: an empirical study from China [D]. Chongqing University, 2012

[2] Zhang Liangjun, Yang Tan, Xiao Gang, Xu Shengbing. MATLAB Data Analysis and Data Mining [M]. China Machine Press, 2015

[3] Feng Chao. Research of K-means clustering algorithm [D]. Dalian University Of Technology, 2007

[4] Jiang Li. The research on fuzzy c-means cluster analysis and its applications[D]. Zhejiang Gongshang University, 2010

[5] Ma Nana. Research on fuzzy clustering analysis algorithm based on density [D]. Inner Mongolia University of Science and Technology, 2012

[6] Li Xiongwei. The research on data mining of improved fuzzy C-means clustering algorithm [D]. Lanzhou University of Technology, 2013

[7] Yu Changjun. Research and implementation on fuzzy C-means algorithm for big data in cloud [D]. Wuhan University Of Technology, 2014

[8] Xu Xiaoyong. Program design of fuzzy classifying-analysis based on Matlab [J]. Journal of
Yunnan Nationalities University (Natural Sciences Edition), 2006(03):196-197+210.