2018 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science

# Particle Swarm Optimization Algorithm Based on Greedy Strategy for Solving MOP Crowd Pricing Problem

Yue Zehua
Wuhan University of Technology,
Wuhan 430063, China
e-mail: hua1318195723@qq.com

***Abstract*—Aiming at the problem of the task location and the crowdsourcing pricing of the task package under the bilateral mechanism, a multi-objective optimization (MOP) model was established, and the particle swarm optimization algorithm based on the task point perspective and the user perspective based greedy strategy was adopted as the fitness function (PSO) to solve. The analysis of the pricing law of the task and the judgment of the reasons for the unfinished task are considered. Firstly, we visualize the task and member's feature information, conduct preliminary analysis, and then further analyze the pricing law through K-means clustering. Secondly, we construct the Logit two-dimensional selection model, and through the function form and regression effect of the model, the impact task is not obtained. The reason for the completion is mainly caused by factors such as insufficient credit and time at the time of pricing**

***Keywords-Greedy strategy; Improved PSO algorithm; Crowdsourcing pricing***

## I. Introduction

In recent years, crowdsourcing has received widespread attention from industry and academia [1]. With the rapid development of the Internet, various emerging concepts of the Internet have emerged, including the concept of “crowdsourcing” [2]. “Crowdsourcing” refers to the practice of corporate agencies outsourcing their work tasks to the mass network community on a voluntary basis[3].It can help the company complete internal control, research and mass data collection through the “crowdsourcing” platform [4]. And users can use the “crowdsourcing” platform to use the idle time to earn rewards [5]. Therefore, the "crowdsourcing" network platform is emerging with a strong momentum. Among them, APP such as “Photographing to make money” is an Internet-based self-service labor crowdsourcing platform [6]. The user is registered as a member of the app by downloading [7]. They receive the tasks they need to take photos from the APP, and earn the rewards that the task is calibrated by completing related tasks. Through the above explanations, we can see that commodity pricing is the key link to maintain the “contractor” and “contractor” of the crowdsourcing platform. When the task pricing is reasonable, the task will be completed smoothly [8].

At present, some researchers have summarized the relevant research work of crowdsourcing from different angles. The progress of crowdsourcing is summarized from four aspects: application, algorithm, performance and data set. Explain the challenges faced by crowdsourcing in 12 aspects: synchronous collaboration, real-time response and motivation…We reviewed the crowdsourcing system applied on the World Wide Web, and classified the crowdsourcing system according to the type of problem and the way of collaboration [9].

## II. Logit regression model

### A. Problem description and analysis

We study the pricing rules of each task point and analyze the reasons for the unfinished tasks based on the data. We consider that the pricing of each task point is not only related to itself, but also related to the members around the mission. Firstly, the task and member's latitude and longitude are visually analyzed on the map, and then the K-means clustering analysis results support the pricing law. We construct the Logit two-dimensional selection model, and through the regression effect of the model, we can conclude the reason why the impact task is not completed.

### B. Model establishment

#### 1)Data preprocessing

The problem requires us to analyze the task point and analyze the pricing rules and task completion of the point. The feature information of the task point also includes information that the member himself or herself may select to select the task point. At the same time, the information carried by the members also determines the pricing of the task point, so we will preprocess the data for both the task point and the member.

a) Coordinate and time conversion

To facilitate the creation and solution of the model, we convert the latitude and longitude to geodetic coordinates and measure the time in seconds.

b) Membership density

After visualizing the members and task locations, we found that the density of members has a certain relationship with the pricing of the task points. Therefore, the membership density around each task point is calculated as an independent variable of post processing.

Therefore, taking the task point as the center, the number of members in the circle is approximately as the member density. deni indicates the membership density around the task i.

c) The degree of selection of a task in the area

We believe that for a single variable, the higher the predetermined limit of a member, the higher the probability that the task will be selected by the member, so we use the average scheduled task limit of all members in the region as the selected degree of the task: $lim_i$.

978-1-5386-7445-1/18/$31.00 ©2018 IEEE
DOI 10.1109/DCABES.2018.00071

$$lim_i = \frac{\sum_{i=1}^{n} y\_lim_i}{n} \quad (1)$$

n is the number of members in the area to which the task belongs.

2) Comment on pricing rules

a)Information visualization

Since the location information of the task is given in the form of latitude and longitude, we import the data into the relevant map software to realize the visualization of the data.

Problem 1 requires us to seek the pricing rules of the task points. Therefore, we use the price as the characteristic information to identify each task point in the map, which is shown in Figure 1：


Figure 1. Price classification task scatter plot

From Figure 1, we can get:

①The general trend of price changes, centered on the lighter color area, the surrounding color gradually deepens, and also indirectly indicates that the price is gradually increasing from low to high.

②According to the geographical location, we can find that the lighter areas are generally distributed in the urban areas of the four larger cities such as Guangzhou, Foshan, Dongguan and Shenzhen.

b)Regular verification

After analyzing Figure 1, we believe that the pricing rules of the mission points are related to the density of members in the mission area, the economic development of the geographic location of the mission point, and the distance from the central city.

For the information extraction of the data table, each member has a corresponding predetermined task quota, abstracted as the degree of selection of each task in the preprocessing, We suspect that the selection of the task is related to the pricing of the task, so the data is analyzed by k-means clustering.

We extract the original data set（$x_1,x_2,...,x_n$）.$x_i$ is a 3-dimensional vector: the membership density, the degree of selection and the pricing of each task that affects the variables.

$$\arg_S min \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - u_i||^2 \quad (2)$$

$u_i$ is the average value of $S_i$ .

After the classification is finished, it is divided into three groups in a visual map, and the progressive color represents the degree of development of the region. as shown in picture 2:

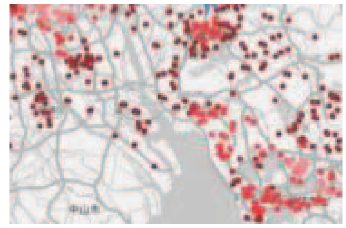

Figure 2. k-means clustering map

After k-means clustering, three types of results are obtained, which are still radiated around Foshan City, Guangzhou City, Dongguan City, and Shenzhen City. In the clustering results, each central city belongs to one category, so our original The analysis is established, and the resulting task pricing rules as follows:

①In areas with high member density, due to the influence of competition, the task pricing is generally low.

②In areas close to the central city, mission pricing is generally low.

③In areas with relatively high levels of economic development, prices are lower.

3) Model establishment and solution

The task completion degree P is a qualitative variable, which we can get from the attachment:

$$p_i = \begin{cases} 1 & \text{Task completed} \\ 0 & \text{Task not completed} \end{cases} \quad i = 1,2...n \quad (3)$$

Since the traditional linear probability model may have problems with fitting values less than 0 or greater than 1, we refer to the relevant literature[1].Introducing latent variables Construct a Logit two-dimensional selection model, As follows:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \quad (4)$$

$u_i$ is disturbance term. Its cumulative distribution is a logistic distribution. The fitted value of the dependent variable represents the logarithm of the likelihood of $P_i$=1.

It can be known from equation (4) that when $P_i$=0 and $P_i$=1 dependent variable is meaningless. Therefore, the maximum likelihood estimation method is used. The data is solved by taking in the preprocessed data to obtain:

$$\ln\left(\frac{p_i}{1-p_i}\right) = -3.382 + 0.045 \mathrm{pri}_i - 0.008 den_i + 0.13 lab_i + 0.064 \lim_i \quad (5)$$

Among them, $pri_i$ is the price of the i-th task. $den_i$ is the area density of the i-th task. $lab_i$ is the area number to which the i-th task belongs. $lim_i$ is the average predetermined limit value for the i-th task.

C. Result analysis

We use the task price, regional density, label, average predetermined limit and average reputation as the explanatory variables of Logit regression, and enter the SPSS statistical software as the dependent variable to obtain the logit model. We find that the more specific one is that the estimate of the average reputation approaches 0, so the explanatory variable is not presented in equation (3). At the same time, the significance level of its coefficient estimates is significant at the saliency level of 5%. [2] Therefore, the null hypothesis is rejected, that is, the average credibility is considered to have no effect on the completion of the task at a level of 5% significance.

At the same time, we get the following grouping appropriate rate table, see Table 1. There are a total of 313 groups attributed to the unfinished group of tasks, of which 69.6% are correctly grouped and 525 groups are attributed to the task completion group, and the correct grouping ratio is 92%.

***Table 1 Grouping appropriate tables***

| act/ret | 0 | 0 | Cor |
|---|---|---|---|
| 0 | 218 | 95 | 69.6 |
| 1 | 42 | 480 | 92.0 |
| Over | | | 81.3 |

From Table 1, we can see that the classification accuracy rate of task completion is 92.0%, and the uncompleted classification appropriate rate is 69.6%, the overall classification appropriate rate reached 81.3%. Explain that the results we get with Logit are acceptable [3], However, some tasks are still not completed, and the reasons are as follows:

(1) It can be seen from the logistic regression analysis table that except for the reputation factor, the other factors determine the degree of completion of the final task to a certain extent, but the degree of influence is also different, in which the member density and the price of the task are completed. The reason for the impact is greater.

(2) Analysis function form We can see that the average reputation factor is not reflected in the completion of the task, but in actual circumstances, there is a high probability that users with high quotas and low reputation will lack the consideration of reputation. To a large extent, the task completion rate is low, Therefore, we have reason to believe that users with low credibility may not complete the task after accepting the task, resulting in the task not being completed.

In summary, we analyze that the main reason for influencing the task is that there is a big shortage of considerations for the user's reputation. At the same time, it may also be due to the density of members and the impact of the price of the task on the completion rate.

.

.

## III. CONCLUSION

This paper mainly builds a model around the crowdsourcing pricing problem. When using k-means cluster analysis, time can be considered on the basis of the original variables, and the geographical relationship between tasks can make the analysis of clusters more reliable and sufficient information. When establishing a multi-objective planning model for task points, the threshold setting of the greedy algorithm can be optimized, especially when selecting members in the area of the task point, the time can be considered, the association among the members, etc., and the variables are selected to be robust. When using the “pso” algorithm to optimize the completeness, you can consider the situation of the local optimal solution in the optimization process, and then set the weight coefficient to adjust the update speed of the particle.

## REFERENCES


[1] Pan Shichu. Econometrics (Fifth Edition) [M]. Beijing: China Renmin University Press, 2015: 197-210.

[2] U.Gadiraju, G.Demartini, R.Kawase and S.Dietze,Human Beyond theMachine: Challenges and Opportunities of Microtask Crowdsourcing, pp.81-85, July-Aug.2015.

[3] Tao Changqi. Econometrics [M]. Nanjing: Nanjing University Press, 2015: 207-213.

[4] Xu Zhikai, Zhang Hongli, Yu Xiangzhan, Zhou Zhigang. IoT search task allocation mechanism based on combined two-way auction [J].Journal of Communications, 2015, 36(12): 47-56.

[5] Zhuo Jinwu. Application of MATLAB in Mathematical Modeling (Second Edition). Beijing: Beijing University of Aeronautics and Astronautics Press, 2014: 150-158.

[6] Wang Zai-Rong, Sparsing algorithm of massive dispersed two-dimensional data based on K-Neighboring, Chengdu, 2010, pp. 376-379.

[7] Alonso, Baeza-Yates R. Design and implementation of relevance assessments using crowdsourcing. Proceedings of the 33rd European Conference on IR Research. On Aggregating Lab Dublin, Ireland, UK, 2011:153-164.

[8] Ipeirotis P G, Provost F, Wang J. Quality management onamazon mechanical turk, Proceedings of the ACM SIGKDDWorkshop on Human Computation. Washington, USA, 2010: 64-67

[9] Blanco R, Halpin H, Herzig D M, et al. Repeatable andreliable search system evaluation using crowdsourcing , Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 923 - 93.