Application of cluster analysis in optimization problem of pricing of a mobile Internet App

Sixuan Bai¹,³, Daiqin Liao², Mei Wang¹ and Yang Ma¹

¹College of Mathematics and Statistics, Chongqing Jiaotong University, Chongqing 400060, China  
²School of Science, Lanzhou Jiaotong University, Lanzhou 730050, China  
³Corresponding author's e-mail: bxsx621@163.com

Abstract. This paper focus on the problem of task pricing and completion of the tasks of a mobile Internet APP. Based on the original data set, the original pricing model of the project's tasks was established by using the multiple linear regression analysis. The influences of factors such as the task point’s pricing, number of other tasks, and total task price on tasks completion were analysed. Further, based on the factors which affecting task completion, a new task pricing model was developed by using cluster analysis to tailor the price of some task points and significantly improved the completion of task without increasing the total task price.

1. Introduction

Crowdsourcing is a new business model with the characteristics of extensive collaboration, multi-party interaction and flexible organization and management. It plays an important role in promoting the optimal allocation and sharing of various innovation resources and innovation factors, and is an important source of entrepreneurial innovation in various fields of production and life in the context of the "Internet+" era.

"Photo money APP" is a kind of self-help crowdsourcing mode under the mobile Internet. Use the APP to register as a member, then receive tasks (such as supermarkets to check the shelves of certain goods), and then completing the task will earn the reward set by the APP for the task. This self-service labor crowdsourcing platform based on mobile Internet not only provides more effective data, shortens the survey time and ensures the authenticity of the survey data, but also saves the cost of the survey to a large extent. The core factor of the platform operation is the task pricing. Reasonable pricing is related to whether someone gets the task, and ultimately affects the completion of the task.

There are many methods to study such problems, such as using multiple linear regression [1] to students scores, cluster analysis [2] to analyze mobile banking adoption, gray correlation analysis to study cloud service trust evaluation [3] and analysis of energy consumption, environmental pollution and economic growth [4], RBF neural network models [5] to determine new pricing scheme and stock price [6]. However, in the existing studies, when considering the influencing variables of the task pricing model, some important factor variables are ignored. When processing the data, the amount of calculation is large and easy to make mistakes. The present models are easy to produce deviation compared with the real data, etc.

In this paper, we used multiple linear regression analysis to model the original pricing law of the project, and analyzed the influence of some important factors, such as the task point’s pricing, number of other tasks, and total task price. Meanwhile, using the independent sample mean test [1], the factors
affecting task non-completion were analyzed with the help of SPSS software. Then, based on the factors that affect the task completion, a new task pricing model is established by using R language for cluster analysis, which greatly reduces the amount of computation and enhances the reliability.

2. Task pricing model of the original data set and the reasons for task non-completion

2.1. Task pricing model of the original data for the project
Since the number of task points of original data set provided in Appendix is too large, but in order to improve the accuracy and enhance the extensiveness of the model, we chose to use all the data provided. And a range of 1.5 km for each task point was selected as the study area for that task point, as shown in Figure 1.

![Figure 1. Range for each task point.](image)

To more fully reflect the factors affecting the task pricing pattern, we selected three representative impact indicators in original data set within the 1.5 km of the task point: the number of task \( x_1 \), the number of members \( x_2 \), and total reputation value of members \( x_3 \). Because: (1) "Quantity precedes price". The denser the distribution of task points within the same range, the more tasks will be completed, which will further affect the task pricing. (2) Within the same range, the denser the number of members is, the greater the competition will be, which will also affect the task pricing to some extent. (3) Since the time and reservation limit of the member are given by reference to its reputation, the total reputation value of the member can fully reflect the reservation limit and the time of the start of the reservation, so the total reputation value of the member within the scope of the unit is also an important factor affecting the pricing law.

Regression analysis is a statistical technique for estimating the relationship among variables which have reason and result relation. Main focus of univariate regression is analyse the relationship between a dependent variable and one independent variable and formulates the linear relation equation between dependent and independent variable. Regression models with one dependent variable and more than one independent variables are called multilinear regression [1]. In this study, based on the above data processing and multilinear regression, the mathematical model is established as follows:

\[
P = a_1 x_1 + a_2 x_2 + a_3 x_3 + b
\]  

(1)

Where \( P \) is the task price, \( a_1, a_2, a_3 \) the weight coefficients of respectively \( x_1, x_2, x_3 \) and \( b \) is a constant that can be approximated by the principal component analysis method [1].

The results obtained by performing multiple linear regression for \( a_1, a_2, a_3, b \) within each task point \( x_1, x_2, x_3 \) are as follows (Table 1).

| Models | Non-standardized coefficient | Standardized coefficient | t   | Sig  |
|--------|-----------------------------|--------------------------|-----|------|
|        | B Standard Error            | Trial Version            |     |      |
| \( b \) | 71.910                      | 0.258                    | 278.973 | 0.000 |
| \( a_1 \) | -0.327                      | 0.061                    | -5.346 | 0.000 |
| \( a_2 \) | -0.308                      | 0.029                    | -10.491 | 0.000 |
| \( a_3 \) | -5.351E-5                   | 0.000                    | -2.558 | 0.011 |

Table 1. Coefficients for Eq. (10) obtained by SPSS.
Obtaining 71.910, -0.327, -0.308, -5.351*10\(^{-5}\) as the values of the weight coefficients, \(a_1, a_2, a_3\), and the value of \(b\) as 71.910, there is obtained a linear relationship between the task price \(P\) and the indicator \(x_1, x_2, x_3\). Thus, the task pricing model of the original data for the project is established.

\[
P = -0.327x_1 - 0.308x_2 - 5.351\times10^{-5}x_3 + 71.910
\]  
(2)

2.2. **Analysis for the reasons of task non-completion**

According the task pricing model of the original data for the project shown in Eq. (2), it is clear that the task pricing, the number of tasks, the number of members, and total reputation value of members affect the task completion. Therefore, SPSS software was used to carry out mean test on the data corresponding to these indicators to analyse for the reasons of task non-completion.

2.2.1. **The effect of task pricing on task completion of a fixed range.** Within a fixed range, SPSS software is used to carry out mean test on the data corresponding to these indicators. Tables 2 and 3 were derived from the tests. According to Table 2, the number of completed tasks is 522 and the mean price is 67.9281, while the number of uncompleted tasks is 313 and the mean price is 69.8199. The number of completed tasks is greater than the number of uncompleted tasks, and the mean price of completed tasks is greater than the mean price of uncompleted tasks. The number of completed tasks is larger than the number of uncompleted tasks, and the average price of completed tasks is larger than the average price of uncompleted tasks, so we can propose the hypothesis that the uncompleted tasks are related to the low pricing of tasks. In Table 3, F is variance test, df is short for degree of freedom, t means T-test result, Sig represents the significance. The Sig.(two-sided) < 0.05 for the number of tasks indicates that task completion is related to the pricing of tasks within a fixed range, and the higher the task pricing the higher the corresponding task completion, and the lower the task pricing the lower the corresponding task completion.

| Groups   | N   | Mean value | Standard deviation | Standard error of the mean |
|----------|-----|------------|--------------------|---------------------------|
| Uncompleted | 313 | 67.9281    | 3.66466            | 0.20714                   |
| Completed  | 522 | 69.8199    | 4.81801            | 0.21088                   |

**Table 2.** Group statistics of price.

**Table 3.** Independent sample test of price.

|                      | Levene's test for variance of equations | t-test for the variance of the mean | 95% confidence interval of the difference |
|----------------------|----------------------------------------|------------------------------------|------------------------------------------|
|                      | F | Sig | t    | df | Sig (two-sided) | Difference of means | Standard error value | Lower limit | Upper limit |
| Assuming equal variances | 15.511 | 0.000 | -5.985 | 833 | 0.000 | -1.89181 | 0.31608 | -2.51221 | -1.27140 |
| Assuming unequal variances | -6.400 | 0.000 | 787.376 | 0.000 | -1.83181 | 0.29560 | -2.47206 | -1.31156 |

2.2.2. **The effect of the number of tasks on task completion under the same pricing in a fixed range.** According to Table 4, it can be seen that the number of task completions near the control task with the same pricing of $65 is larger than the number of tasks uncompleted, so we can hypothesize that task
incompletes are related to the number of tasks in the fixed range. According to Table 5, the Sig.(two-sided) of the number of tasks < 0.2 indicates the task completion is related to the number of tasks in the fixed range with the same task pricing, the more the number of tasks corresponds to the higher task completion, and the less the number of tasks corresponds to the lower task completion.

Table 4. Group statistics of the number of tasks.

| Completion | N  | Mean value | Standard deviation | Standard error of the mean |
|------------|----|------------|--------------------|---------------------------|
| 0.00       | 30 | 5.5000     | 1.88917            | 0.34491                   |
| 1.00       | 50 | 6.4000     | 1.89737            | 0.32071                   |

Table 5. Independent sample tests of the number of tasks.

| The Levene test for equation variance | t-test for the mean equation | 80% confidence interval of the difference |
|---------------------------------------|-----------------------------|------------------------------------------|
| F | Sig | t   | df | Sig (two-sided) | Mean Variance | Standard error value | Lower limit | Upper limit |
| 0.534 | 0.468 | -1.910 | 63 | 0.061 | -0.90000 | 0.47114 | -1.51019 | -0.28981 |
| Assuming equal variances | | | | | | | | |
| Assuming unequal variances | -1.911 | 61.569 | 0.061 | -0.90000 | 0.47098 | -1.51013 | -0.28987 |

2.2.3. The effect of the number of members on task completion under the same pricing in a fixed range. According to Table 6, we can see that the number of unfinished tasks is less than the number of completed tasks under the same pricing of tasks in the fixed range, and the mean value of the number of members in the fixed range of task points with unfinished tasks is higher than the mean value of the number of members in the fixed range of task points with completed tasks, so we can hypothesize that the unfinished tasks are related to the number of members in the fixed range. In Table 7, the Sig.(two-sided) < 0.2, thus, the task completion is related to the number of members in the fixed range with the same task pricing, the higher the number of members the corresponding task completion is relatively lower, the lower the number of members the corresponding task completion is relatively higher.

Table 6. Group statistics of the number of members.

| Completion | N  | Mean value | Standard deviation | Standard error of the mean |
|------------|----|------------|--------------------|---------------------------|
| 0.00       | 30 | 93.8667    | 52.40475           | 9.56775                   |
| 1.00       | 35 | 68.7429    | 72.79679           | 12.30490                  |
Table 7. Independent sample test of the number of members.

|                          | The Levene test for equation variance | The levene test for equation variance | 80% confidence interval of the difference |
|--------------------------|---------------------------------------|---------------------------------------|------------------------------------------|
|                          | f  | Sig | t   | df | Sig(two-sided) | Mean Variance | Standard error of the mean | Lower limit | Upper limit |
| Assuming equal variances | 2.149 | 0.148 | 1.572 | 63 | 0.121          | 25.12381      | 15.91821           | 4.42988 | 45.81774 |
| Assuming unequal variances | 1.612 | 61.279 | 0.112 | 25.12381 | 15.58694 | 4.93062 | 45.31700 |

3. The new task pricing scheme

3.1. New task pricing scheme model

The total number of tasks within a fixed range around the task point $x_1$, the number of members within a fixed range around the task point $x_2$, and the total credibility of members within a fixed range of the task point $x_3$ have been defined above. According to data of APP platform, the completion of tasks put on the APP platform is not optimistic, and we find the task pricing $P$ is the main influencing factor of the task situation in the analysis of the reasons for the task not completed in section 2. Therefore, we design a new task pricing scheme for the projects in appendix, and set the task pricing at $P$ this time. Then, the total number of tasks in the fixed range of task points $x_1$, the number of members in the fixed range of task points $x_2$, and the total reputation of members in the fixed range of task points $x_3$ remain unchanged. Clustering methods form groups or clusters of objects based on some measure of similarity of attributes between those objects. In a typical Operations Management application, these objects can be manufacturing facilities, business units, firms, or managers; while the attributes are variables such as firm design flexibility, business size, or tenure of managerial experience. Similarity between these objects is commonly expressed in a measure of proximity represented quantitatively by the squared Euclidean distances between pairs of objects based on the set of variables obtained from single- or multi-item scales [2]. Therefore, following this idea, the new task pricing scheme is modelled as follows:

(1) Through the analysis of the reasons for the task not completed in question one, we get the task pricing $P$ as the main factor affecting the task situation, and through the analysis of the task pricing law we can know that these three indicators $x_1, x_2, x_3$ are the most direct factors affecting the task pricing. Therefore, in this paper, we use R language in building the model to cluster the data provided in appendix as the classification basis for cluster analysis. According to the task pricing provided in Annex 1, we analysed the change pattern. Since the price range is (65,85) and the price difference between each pricing at (65,75) is 0.5, it can be divided into 22 price classes. Based on this, the prices of 835 task points were classified into 22 price classes. Therefore, all task points were grouped into 22 categories according to different price levels in this paper. On this basis, after using R language to cluster them, it is found that there are obvious differences in prices between each category, so this clustering is reasonable. $x_1, x_2, x_3$ of each class are similar, so its popularities among members are similar. Therefore, this paper gives a unified price for each task in the same class, that is, sum all prices in the class and take the average value.
(2) We set a uniform pricing for each of the 22 classes \( P_j \ (j = 1, 2, 3, \ldots, 22) \), with a task point \( n_j \) in each class and a pricing for each task point \( P_j \). Then,

\[
P_j = \frac{1}{n_j} \sum_{j=1}^{n_j} P_j \quad (j = 1, 2, 3, \ldots, 22)
\]

We used Excel to calculate the pricing of task points in the same class, and thus figured out that all 835 task points had new pricing \( P_m \ (m = 1, 2, 3, \ldots, 835) \).

(3) Because support vector machines (SVMs) are widely used in models for forecasting values [7]. A support vector machine classifier is trained according to the four indicators \( x_1, x_2, x_3, p \) for the completion of 835 task points in Section 2. Then, it is used to classify the \( x_1, x_2, x_3, P_m \) given by the new scheme to obtain the completion of task points. By comparing the completion number of task points obtained by the new scheme with that obtained by the original scheme, it can be concluded whether the new scheme is useful.

### 3.2. Analysis and comparison of calculation results

In order to verify that there is a difference in the new pricing between the groups of each class, i.e., there is rationality. We used SPSS software to test the significance of the mean value of the 22 categories clustered above, and the results are shown in Table 8. There are some differences in pricing among the 22 categories of income. According to Table 8, the significance of inter-group pricing of each class is less than 0.05, so there is a significant difference in inter-group pricing of each class, so the classification is reasonable.

|                        | Sum of squares | df | mean square | F     | Significance |
|------------------------|----------------|----|-------------|-------|--------------|
| intergroup             | 6549.359       | 21 | 311.874     | 24.298| 0.000        |
| Intra-group            | 10435.144      | 813| 12.835      |       |              |
| Total                  | 16984.503      | 834|             |       |              |

The completion of the 835 task points in Section 2 was trained by a support vector machine classifier based on four indicators \( x_1, x_2, x_3, P_j \). According to the calculation, we can conclude that the number of tasks completed under the original task pricing is 522, and the number of tasks completed after the implementation of the new scheme through the clustering-adjusted pricing and the use of SVM method under the same total price is 629. That is to say, the task completion rate has increased by 20.5% compared with the original scheme given by the project in Appendix. Thus, the pricing adjustment here of the new scheme is successful.

### 4. Summary

This paper analyses the task pricing and completion of "photo money APP " under the mobile Internet. According to the original data set, in order to better analysed the task pricing law, we used three indicators that describe the popularity and ease of execution of tasks: the number of tasks, the number of members, and total reputation value of members. Then, a multiple linear regression analysis is used to establish the task pricing model. Also using independent sample mean test method, the factors affecting the task unfinished situation were analysed with the help of SPSS software, and from the results, it was analysed that there were significant differences in the pricing of task points under different task completion situations, the number of other tasks near the task point, and the total reputation value near the task point. Based on the factors affecting task completion, we developed a better pricing scheme: without increasing the total task price, we adjusted the pricing of some task points according to local conditions and significantly improved the task completion. All tasks were clustered according to the three main indicators describing the characteristics of the task points, and then the average price was derived for each subcategory of tasks as the new price for each task in this category. Based on the
original dataset, a classifier was trained based on the support vector machine algorithm to classify and predict the task completion after the new pricing. Finally, the task completion performance of the original pricing scheme and the new pricing scheme were compared separately, and it was found that the new task pricing scheme significantly improved the quality of task completion.

References

[1] Uyanik G K, Guler N 2013 A study on multiple linear regression analysis. *Procedia Soc. Behav. Sci.* **106** 234

[2] Chawla D, Joshi H 2017 Consumer perspectives about mobile banking adoption in India—a cluster analysis. *Int. J. Bank Mark.* **35** 616

[3] Wang Y, Wen J, Wang X, Tao B, Zhou W 2019 A cloud service trust evaluation model based on combining weights and gray correlation analysis. *Secu. Commu. Netw.* **2019** 2437062

[4] Luo X, Lu L, Wang Z, Yang L 2019 Gray correlation analysis of energy consumption, environmental pollution, and economic growth in subtropical regions of China: Guangxi and Zhejiang as examples. *Trop. Conserv. Sci.* **12** 1

[5] Zhang Y, Zhou Q, Sun C, Lei S, Liu Y, Song Y 2008 RBF neural network and ANFIS-based short-term load forecasting approach in real-time price environment. *IEEE T. Power Syst.* **23** 853

[6] Song Y G, Zhou Y L, Han R J 2018 Neural networks for stock price prediction. *arXiv preprint arXiv:1805.11317.*

[7] Vilela L F, Leme R C, Pinheiro C A, Carpinteiro O A 2019 Forecasting financial series using clustering methods and support vector regression. *Arti. Intell. Rev.* **52** 743

Appendix

Original data used in this paper can be find in the following link with fetch code h63a: https://pan.baidu.com/s/1MkxOyDxbxEGr_WsH5IMFA