文章编号：1004-4353(2018)02-0164-06

基于贪婪算法的众包平台定价规律的研究

张月蕾， 崔连标， 朱家明*
（安徽财经大学 统计与应用数学学院，安徽 蚌埠 233030）

摘要：针对“拍照赚钱”任务定价，利用数据可视化、K-means 聚类等方法，分析定价规律和任务未完成的原因. 对任务完成度低的区域，使用贪婪算法优化任务分配，并重新调整定价方案. 最后，以广州市为例进行线性定价，验证了调整后定价方案的合理性. 本文模型具有较好的可行性和有效性，能够为一般的定价机制提供参考.

关键词：任务定价；可视化；K-means 聚类分析；贪婪算法；回归拟合
中图分类号：F224 **文献标识码**：A

Research on the law of crowdsourcing platform pricing based on greedy algorithm

ZHANG Yuelei， CUI Lianbiao， ZHU Jiaming*
（*School of Statistics and Applied Mathematics*, *Anhui University of Finance and Economics*, *Bengbu* 233030, *China*）

Abstract: Aim at pricing of “photo-making money” tasks, using data visualization, K-means clustering and so on etc. to analyze pricing rules and the reasons for unfinished tasks. In areas with low mission completion, greedy algorithms are used to optimize task allocation and readjust pricing plans. Finally, the linear pricing in Guangzhou is taken as an example to verify the rationality of the adjusted pricing plan. The results show that this model has good feasibility and effectiveness and can provide guidance for the general pricing mechanism.

Keywords: task pricing; visualization; K-means clustering; greedy algorithm; regression fitting

“拍照赚钱”是基于移动互联网的自助式劳务众包平台，可为企业信息搜集和各种商业检查提供自助式服务. 相比传统的市场调查方式，这种自助式调查不仅可大大节省调查成本，有助于保证调查数据的真实性，也可随时掌握项目的动态信息. 平台任务能否完成的关键因素是任务定价，即只有任务定价合理才能保证任务完成的效率，否则只能造成任务无人问津，最终导致任务的失败.

目前，一些学者对定价方法进行了研究，并取得了一些研究成果. 例如：张鹏[1]基于委托代理的众包机制，对公司自营和第三方运营两种模式提出了线性定价机制，发现线性奖金的激励机制优于固定奖金的激励机制；但是该结论并未考虑云端与参与者的筛选情况，因此存在一定的缺陷. Y. Singer 等[2]提出了一种在线定价机制，综合考虑了企业与会员两方面因素；这种在线定价策略虽然相对合理，但需要多次调整任务定价，增加了任务选择时间. 刘晓钢[3]对威客网的定价数据进行了分析，指出任务定价受自身属性和市场竞争状况的影响，并结合实证对所提假设进行了验证；但是，该研究仅对一个最大的众包网站进行了分析，缺乏与其他典型众包网站的对比分析. 基于上述研究成果，本文考虑任务自身属性与会员两方面因素，尝试结合贪婪算法对参与者进行筛选，优化任务分配，以期确定最优定价方案.

收稿日期：2017－12－21 * **通信作者**：朱家明(1973—)，男，副教授，研究方向为统计与应用数学.
基金项目：国家自然科学基金资助项目(61703001)；安徽财经大学大学生创客实验室项目(acdxsck201707)

1　数据来源与模型假设

研究数据取自 2017 年“高教社杯”全国大学生数学建模竞赛 B 题. 数据涵盖：一个已结束的任务数据，包括每个任务点定位（经纬度）、定价和完成情况；会员信息数据，包括会员定位（经纬度）、信誉值、会员开始预订时间和预订限额；一个包括任务位置信息的新项目数据. 为便于问题分析，本文提出以下假设：①会员选择任务时是理性的，优先选择距离近、价格高的任务；②不同任务的完成难度相同，即各个任务的完成时间不存在显著性差异；③会员位置表示自身定位，而不是固定的居住 IP 定位；④任务的单位距离成本不存在差异.

2　基于 K-means 聚类的任务定价分析

2.1　数据处理

以已经结束的项目数据为样本，主要考虑任务属性和会员信息，并分析定价策略. 将会员位置、任务数据进行可视化，如图 1 所示. 其中，点的大小对应定价的高低；点的亮度表示任务完成的情况，最亮实心圆代表未完成任务，灰色实心圆代表已完成任务，其他深灰色覆盖区域表示会员分布.

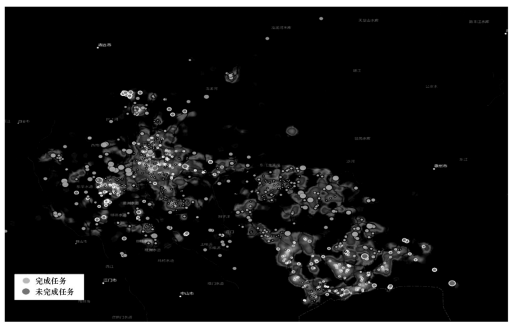

图 1　任务完成情况热力图

由图 1 可以看出：①不同区域任务点分布、定价存在较大差异. ②相对分散的任务点，定价相对偏高；任务点分布密集的区域，定价相对偏低. ③区域任务定价低则任务完成度相对低，定价高则任务完成度相对高.

2.2　结果分析

针对不同区域任务点分布、定价存在差异的问题，对任务位置进行 K-means 聚类分析[4]. 利用 MATLAB 聚类程序，将任务点分成 4 个区域（图 2），4 个聚类中心点的坐标为：(23.014 9°N，113.184 6°E)、(22.663 1°N，114.046 4°E)、(23.278 2°N，113.326 5°E)和(22.956 1°N，113.749 1°E). 计算任务点到类中心的平面距离[5]，再结合任务的标价，可得到任务定价和距离间的相关系数：$\rho_1=0.36$，$\rho_2=0.53$，$\rho_3=0.15$，$\rho_4=-0.41$. 由此可知，任务定价和距离存在相关关系，但相关性在不同区域间存在异质性，这可能与所在区域的经济发展、交通、气候等因素相关，因为这些因素会影响任务完成的成本.

为分析会员聚集程度对任务完成度的影响，对任务信息、会员分布进行可视化处理，如图 3 所示. 图 3 中正方形格网大小相同，每个正方形右下角标签表示该区域内的会员数，左下角和左上角分别对应该

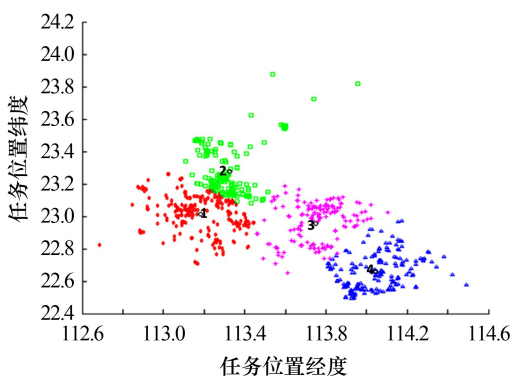

图 2　聚类结果图

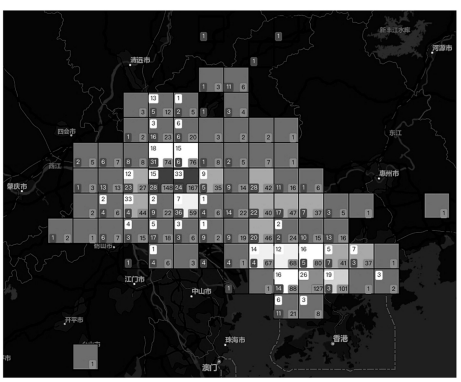

图 3　会员信息网格化

区域完成和未完成的任务总数.统计分割区域中的会员总数,可求得会员分布密度.图 3 显示,任务完成度达到 100%的网格占 60.98%,剩余 39.02%的网格中任务未被全部执行.这可能与会员密度相关,因为和此处会员密集,任务点会员分布不均衡.

3 基于贪婪算法的任务定价分析

3.1 参考样本选取

经统计,已经结束的 835 个任务分布在 4 个地级市,涵盖了 25 个区域,各区域任务完成情况见表 1.由表 1 可知,佛山市南海区、广州市南沙区、广东省东莞市等 8 个地区任务完成度达到 92%以上,说明这些地区定价结构相对合理,可作为定价参考样本.对剩余 17 个区域进行重新定价,并与参考样本进行比较,以此评估定价策略的有效性.

表 1 不同区域任务完成度

区域	任务完成度/%	区域	任务完成度/%	区域	任务完成度/%
深圳市福田区	0.00	广州市萝岗区	50.00	佛山市南海区	92.75
深圳市盐田区	0.00	广州市荔湾区	54.55	广州市南沙区	93.75
深圳市罗湖区	10.00	深圳市南山区	55.56	东莞市	100.00
深圳市宝安区	12.86	广州市白云区	56.41	佛山市高明区	100.00
深圳市龙岗区	13.51	佛山市顺德区	57.45	佛山市三水区	100.00
佛山市禅城区	15.00	广州市花都区	57.50	广州市从化市	100.00
广州市天河区	27.03	广州市海珠区	60.00	广州市增城区	100.00
广州市越秀区	35.71	广州市番禺区	78.33	清远市佛冈县	100.00
广州市黄埔区	38.46	—	—	—	—

3.2 模型的建立和结果分析

将会员任务分配问题转化为有界背包问题[6],利用有界贪婪算法[7]求解如下优化问题:

$$\begin{aligned} &\max \sum_{i=1}^{N} \hat{\mu}_i x_i^p, \\ &\text{s.t.} \sum_{i=1}^{N} c_i x_i^p \leqslant (1-\varepsilon)B. \end{aligned} \tag{1}$$

其中:c_i 表示每次分配给会员 i 所产生的成本;μ_i 表示会员 i 每次完成分配任务所对应的回报;x_i^p 是决策变量,表示在利用阶段分配给会员 i 的任务个数;$\hat{\mu}_i = f(c_i)$,表示探索阶段 μ_i 的估计值;B 是总任务预算,用 εB 表示探索阶段部分预算值;$\hat{\mu}_i / c_i$ 表示会员 i 单位成本完成任务的质量(会员信誉值).

将会员按照信誉值进行降序处理.第一轮尽可能选择信誉值最高、优势最大的会员进行任务分配,直至达到所选会员的最大预定任务限额;第二轮选择剩余会员中信誉值最高的会员进行任务分配,直至达到所选会员的最大预定任务限额;重复以上筛选步骤,直至会员或者任务无法继续匹配.

模型求解过程包含两个阶段:贪婪算法探索阶段和贪婪算法利用阶段.其中,探索阶段共包含如下 7 个步骤:

步骤 1 选取区域内部分会员,取得“任务预算”(即某个会员能够完成的最大任务个数及位置范围)$B^p = \varepsilon B$,设置 $t = 1$,进行会员任务分配;

步骤 2 令 $B^p = B^p - \sum_{k=1}^{N} c_k$, $t = t + 1$,条件为 $1 \leqslant t \leqslant \lfloor B^p / \sum_{i=1}^{N} c_i \rfloor$,开始当前循环;

步骤 3 对会员按照 c_i 进行升序排序,根据 $B^p < \min c_i$ 判断剩余“任务预算”,如果成立,跳出循环;否则,继续返回步骤 2,选择剩余会员中符合条件的会员;

步骤 4　直至“任务预算”$B^p < \min c_i$，结束当前循环；

步骤 5　根据循环结果，对会员已完成的任务进行质量估值，估值函数 $\mu_i = f(c_i)$；

步骤 6　利用步骤 5 中的函数得到剩余会员的任务质量估值；

步骤 7　算法结束.

贪婪算法利用阶段的设计步骤如下：

步骤 1　根据探索阶段得到的(c_i, μ_i)样本，对每个样本计算单位成本质量 $\tilde{\mu}_i/c_i$，对所有会员按照单位成本质量进行降序排序；

步骤 2　如果 $B^p \geqslant c$（c 表示单位成本质量最高的会员成本），继续执行 $B^p = B^p - c$；否则，将单位成本价值最高的会员去除，直到超出“任务预算”；

步骤 3　结束当前循环，算法结束.

上述贪婪算法程序采用 C++编写. 探索阶段主要是对部分会员进行任务分配，并通过估值函数计算出剩余会员的任务分配；利用阶段主要是对探索阶段的结果进行优化，以此得出区域会员与任务的最优分配.

验证以深圳福田区、广州增城区、南沙区和佛山市三水区为例进行. 利用上述算法，分别得到各区域的任务执行表，见表 2 和表 3. 将这 4 个区域未执行的任务数据与会员分布进行可视化处理，结果见图 4 和图 5. 图 4 中实心圆表示未完成的任务分布，五角星表示会员位置. 图 5 中深黑色点表示未完成的任务分布，灰色点表示会员位置.

表 2　深圳福田区的任务执行情况

任务号码	纬度/°N	经度/°E	区域	任务标价/元	任务执行情况
A0468	22.538 26	114.065 9	深圳市福田区	80	未完成
A0465	22.546 6	114.105	深圳市福田区	67	未完成
A0464	22.527 13	114.053 9	深圳市福田区	65.5	未完成
A0449	22.534 3	114.031 7	深圳市福田区	66	未完成
A0448	22.549 38	114.050 2	深圳市福田区	67	未完成
A0369	22.546 05	114.026	深圳市福田区	66.5	未完成
A0074	22.533 18	114.083 1	深圳市福田区	66.5	未完成
A0072	22.523 31	114.046 6	深圳市福田区	66.5	未完成
A0024	22.542 21	114.019 6	深圳市福田区	66.5	未完成

表 3　广州增城区、南沙区和佛山市三水区的任务执行情况

任务号码	纬度/°N	经度/°E	区域	任务标价/元	任务执行情况
A0126	23.167 69	113.665 2	广州市增城区	75	已完成
A0133	23.186 12	113.597 5	广州市增城区	75	已完成
A0178	22.750 34	113.583 5	广州市南沙区	70	已完成
A0198	22.829 26	113.512 8	广州市南沙区	75	已完成
A0199	22.774 03	113.563 5	广州市南沙区	74.5	已完成
⋮	⋮	⋮	⋮	⋮	⋮
A0737	23.260 48	113.023 6	佛山市三水区	75	已完成
A0745	23.221 14	112.924 8	佛山市三水区	75	已完成
A0829	23.179 03	112.876 2	佛山市三水区	80	已完成

由表 2 可以看出，深圳福田区域的任务标价较低，任务执行度为零，然而该区域内的会员数量较多，会员位置和任务距离并不太大(图 4). 会员不选择接单完成任务，很可能是由于价格偏低所致，因此需要合理地提高任务价格以促使任务完成.

由表 3 可以看出,广州增城区、南沙区和佛山市三水区的任务定价均较高,任务执行度也较高,仅少部分任务未被执行(图 5),未被执行的任务点相对分散且距离会员较远.会员不选择接单完成任务的原因很可能是由于完成任务成本相对较高,因此需要合理地提高定价以促使该部分任务完成.

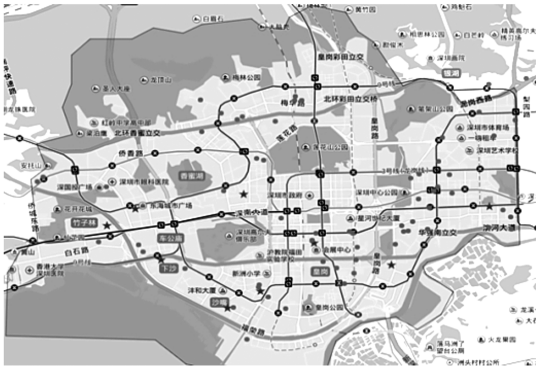

图 4 深圳福田区的会员与未完成任务分布

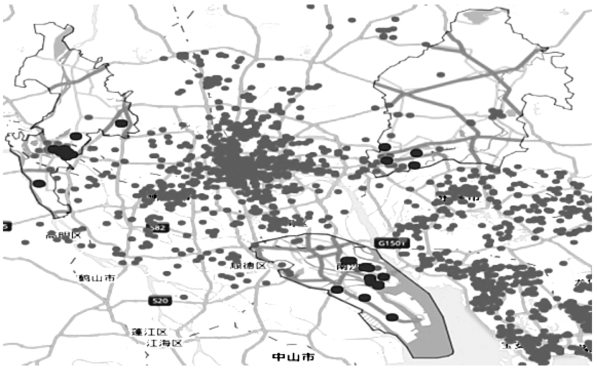

图 5 广州三水区的会员与未完成任务分布

综合考虑会员位置和任务点的距离,再对照 8 个完成度高的区域任务定价机制,本文将剩余 17 个地区的任务定价范围调整为 70～75 元.

4 基于回归拟合的任务定价分析

4.1 模型的建立

基于影响任务定价的因素(任务经度、纬度、会员密度和会员信誉度),以广州市项目数据为分析样本,采用逐步回归的方法[9]建立如下回归拟合模型:

$$y = 2\,563.35 - 20.60C - 6.66V - 22.17M + \xi. \tag{2}$$

其中 y 表示任务定价,C 表示经度,V 表示纬度,M 表示会员密度,ξ 表示随机干扰项.各自变量 P 值均小于 0.05,表明各变量显著.模型中,$F(3,125) = 13.45$,说明模型显著;$R^2 = 0.88$,说明模型拟合效果较好.

利用残差图验证回归模型的合理性,如图 6 所示.由图 6 可以看出,残差项大都分布在 0～3 中,且大致服从均匀分布,说明模型的误差可控.

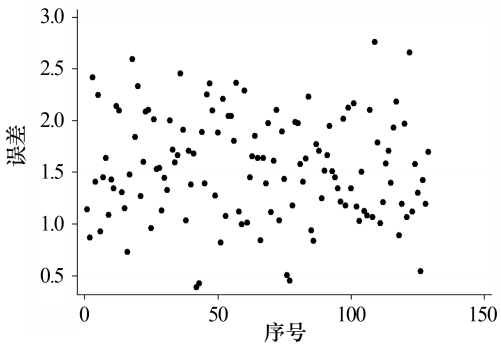

图 6 残差图

4.2 模型的改进

考虑会员信誉度对定价存在的影响,对式(2) 进行改进.由于会员所要完成的具体任务未知,因此不能直接将会员信誉值引入回归模型中,而需要生成一个虚拟变量 D_i[10].采用等距分组的方法,将会员信誉值分段,可得到会员信誉值区间的划分结果:[0,16 999.35]、(16 999.35,33 998.69]、(33 998.69,50 998.04]、(50 998.04,$+\infty$).对应的虚拟变量 D_i 如下:

$$D_i=\begin{cases}0.75, b_i \in [0,16\,999.35];\\ 1.50, b_i \in (16\,999.35,33\,998.69];\\ 2.25, b_i \in (33\,998.69,50\,998.04];\\ 3, b_i \in (50\,998.04,+\infty).\end{cases} \tag{3}$$

其中 b_i 表示会员的信誉水平值. 改进后的回归函数为

$$y=2\,563.35-20.60C-6.66V-22.17M+D_i. \tag{4}$$

选取广州市新项目任务数据为实验样本，验证改进后模型的拟合效果. 将数据代入式(4)计算该项目的任务定价，结果表明其定价范围均在 70～75 元内波动. 该结果与本文在 3.2 中提出的定价区间一致，表明本文模型具有一定的合理性和实用性.

5　结束语

本研究通过对众包平台的定价规律进行分析，得出以下结论：①任务点位置和会员密度对任务定价的影响较大. ②任务定价、任务点分布和会员分布对任务完成度的影响较大. ③本文提出的模型具有合理性，可为一般的众包任务定价提供参考. 由于本文研究搜集对象数据较为困难，仅针对部分区域的定价规律进行了研究，今后将借助大数据挖掘技术丰富研究样本，以完善本文的定价机制.

参考文献：

[1] 张鹏. 基于委托代理的众包式创新激励机制研究[D]. 成都：电子科技大学，2012：4-25.
[2] Singer Y, Mittal M. Pricing mechanisms for crowdsourcing markets[C]//International Conference on World Wide Web. Seoul: ACM, 2013:1157-1166.
[3] 刘晓钢. 众包中任务发布者出价行为的影响因素研究[D]. 重庆：重庆大学，2012：3-21.
[4] 杨娟，屈传慧. 改进 K 均值聚类算法[J]. 舰船电子对抗，2017，40(6)：91-93.
[5] 郭充. 面向 CGCS2000 的格网坐标转换方法及应用研究[D]. 河南：解放军信息工程大学，2009：20-39.
[6] 晏杰. 基于改进的贪婪算法在 0/1 背包问题中的研究与应用[J]. 廊坊师范学院学报(自然科学版)，2011，11(5)：27-30.
[7] 杨子兰，朱娟萍，李睿. 资源受限最小赋权树形图的一种贪婪分解启发式算法[J]. 西南师范大学学报(自然科学版)，2017，42(8)：18-24.
[8] 蒋力，武坤. 0-1 背包问题贪婪算法应用研究[J]. 计算机与数字工程，2007，38(6)：32-33.
[9] 连玉君，杨柳. Stata 中因子变量的使用方法[J]. 郑州航空工业管理学院学报，2018，36(2)：90-103.
[10] 金兰. 基于虚拟变量的分段回归探寻相对稳定点[J]. 延边大学学报(自然科学版)，2005，31(3)：157-160.