APP 拍照任务的定价

摘要

本文建立 APP 拍照任务的定价模型和算法。

对于问题一，以每一个任务为主体，提取该任务的位置，周边会员的位置，预选时间，预选额度以及信誉度信息，将已给的定价视为精准定价，进行线性回归，得出一组系数最优解。为了定性分析任务失败的原因，定义实际和理论定价的差值与理论定价的比值为增值比，进一步根据增值比对任务分类，用每类任务增值比的平均值作为任务能否完成的阈值，借此来给出任务失败的原因。

对于问题二，我们试着从多个角度来回答。角度一是利用上一问中求出的阈值，将任务增值比与其比较，对成功的且价格过高的任务，提取出多余的价格，对失败的且价格偏低的任务，增补欠缺的价格，以此实现价格优化；角度二是优化问题一的模型，考虑变量之间的相互影响，重新回归，得出新的理论价值；角度三是利用任务完成情况确定不等式，采用线性规划的知识确定出系数最优解。

对于问题三，经过分析，采用外围收缩技术，每次利用重心转移法选取最外围的点，从该点开始打包，逐步向内部收缩。打包完成后，将每一个任务包视为一个主体，重新确立回归方程，进行线性回归，求出最优解。之后结合前两问，选取评判任务包完成与否的标准，最后统计任务包完成数目，与之前的结果相比较。

对于问题四，先使用第三问的方法，对任务进行分类打包，然后利用问题三中拟合的线性模型对每个任务包定价。

关键字: 线性回归　阈值比较　外围收缩　重心转移法

1 问题重述

“拍照赚钱”是移动互联网飞速发展下，应运而生的一种新型自助式服务模式。用户成为 APP 的会员后，从 APP 上领取需要拍照的任务，赚取 APP 对任务所标定的酬金。这种基于移动互联网的自助式劳务众包平台，为企业提供各种商业检查和信息搜集，相比传统的市场调查方式可以大大节省调查成本，而且有效地保证了调查数据真实性，缩短了调查的周期。因此 APP 成为该平台运行的核心，而 APP 中的任务定价又是其核心要素。如果定价过高，将导致投资成本的浪费；如果定价过低，有的任务就会无人问津，而导致商品检查的失败。请讨论以下问题：

① 根据已给的信息，包括任务的位置，完成情况，会员的位置，预选时间，预选限额，信誉度，探索定价的规律。

②优化定价方案，重新给出定价，并与原方案比较。

③考虑如何将任务分类聚合，打包分派，以避免恶性竞争，提高运行效率。

④根据以上的模型，对于新一批的任务，给出定价。

2 问题分析

定价策略，市场营销组合中一个十分关键的组成部分。价格通常是影响交易成败的重要因素，同时又是市场营销组合中最难以确定的因素。由此可见定价问题本身就具有一定的难度。而这种新型的自助服务运营模式又有别于以往的定价模型。在问题当中，会员是信息的提供者，即通常意义下的生厂商，而 APP 平台花钱买信息，是消费者。所以在拍照任务定价中，是消费者定价，来吸引生产者，与以往的定价形式有很大的不同，这也正说明了拍照任务定价本身具有一定的特殊性。

给任务定价时通过对任务以及会员信息的筛选，整合，提取，拟合来实现对任务定价规律的探索，从而在此基础上优化定价，包括以下三个步骤：

(1) 任务和会员信息的预处理，包括数据清洗，数据分类，数据整合；

(2) 任务定价规律的探索，应用适当的回归算法找出任务定价与其他信息之间的关系；

(3) 优化任务定价模型，给出更为合理的定价；

(4) 运用已经得到的模型，给新一批的任务定价。

其中步骤 (2)(3) 是算法实现的核心，直接关系到任务定价的合理性，从而影响任务的完成与否。

3 模型假设

(1) 假设会员的任务预定时间只是会员获取任务时竞争力的衡量指标，而不去考察具体的时间分配；

(2) 假设会员的任务限额只是会员获取任务时竞争力的衡量指标，而不去考察会员具体获得多少任务；

(3) 假设任务一经会员领取就会被执行，即视为成功的任务。

4 符号说明

符号	描述
r_i^j	第 i 个任务的经度
r_i^w	第 i 个任务的纬度
$\overline{r^j}$	任务的平均经度
$\overline{r^w}$	任务的平均纬度
R_i^j	第 i 个任务包的经度
R_i^w	第 i 个任务包的纬度
$(\overline{r^j}, \overline{r^w})$	任务的重心位置
h_i^j	第 i 个会员的经度
h_i^w	第 i 个会员的纬度
$\overline{h^j}$	会员的平均经度
$\overline{h^w}$	会员的平均纬度
$(\overline{h^j}, \overline{h^w})$	会员的重心位置
t_i	第 i 个会员的时间
e_i	第 i 个会员的限额
x_i	第 i 个会员的信誉
v_i^s	第 i 个任务的实际定价
v_i^t	第 i 个任务的理论定价
V_i^s	第 i 个任务包的实际价格
V_i^t	第 i 个任务包的理论价格
d_i	第 i 个任务的增值比 =（实际价格-理论价格）/理论价格
D_i	第 i 个任务包的增值比 =（任务包实际价格-理论价格）/理论价格
m_h	平均任务竞争会员数 = 会员预定总限额/任务总数
m_r	平均竞争任务数 = 会员预定总限额/会员总数
b_I	第一类任务（$d_i > 0$，完成）的平均增值比
b_{II}	第二类任务（$d_i > 0$，失败）的平均增值比
b_{III}	第三类任务（$d_i < 0$，完成）的平均增值比
b_{IV}	第四类任务（$d_i < 0$，失败）的平均增值比

5 模型的建立和求解

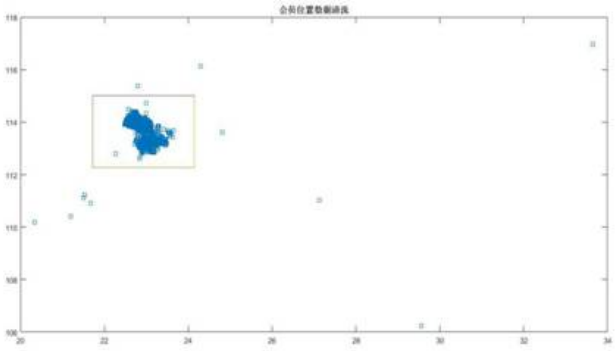
5.1 问题 1 分析

该问要求我们探索任务定价规律。参考附件一，附件二所给信息，我们可知，可能和定价相关的变量有任务的位置，会员的位置，预选限额，预先时间以及信誉度。根据金融经济学中的线性定价法则 [1] 我们假设任务价格和所涉及的变量成线性关系。之后我们选取任务理论价格和实际价格之差的绝对值之和（以下简称绝对值和）作为量度标准，也即绝对值和越小，匹配度越高，模型越发吻合实际。利用线性回归求解出匹配度最高的解。并且根据相关类似模型 [2] 可知，会员和任务的相对距离越远，会员的信誉度越高，预定限额越大，预定时间越短，会员执行任务所需的酬金越高。得出最优解后，需判断是否符合上述规律。接着我们定义增值比=(实际价格-理论价格)/理论价格，用以判断和度量任务实际定价相比于理论定价的变化情况，并以此来求出任务完成或失败的阈值，进而定量地分析任务失败的原因

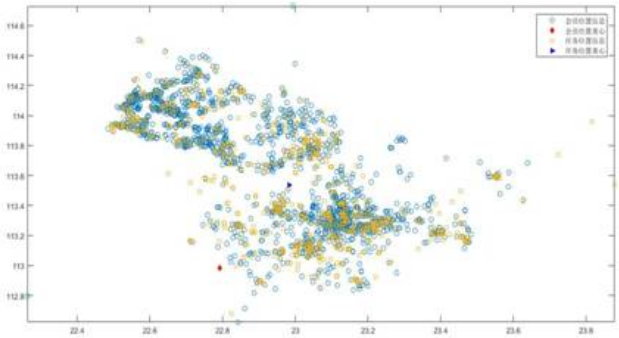
1）问题 1 的模型建立与求解

问题 1 要求对于附件一中的任务探索其定价规律。本文通过以下几个步骤建立模型

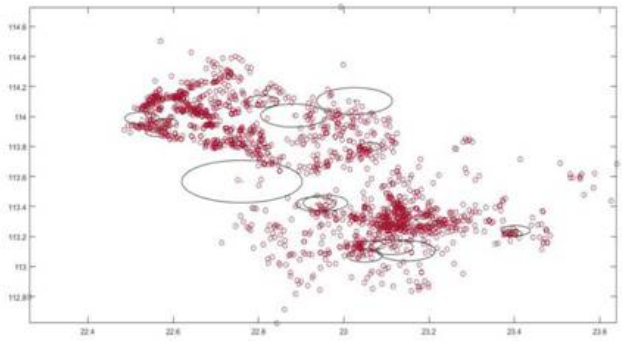
步骤一：首先进行数据清洗，使用统计学的方法，即正态分布 3σ 原则，对任务的位置，会员的位置，预选限额，预先时间以及信誉度进行正态拟合，剔除边缘数据。以会员的位置为例：



然后根据任务和会员的位置信息，利用 MATLAB 画出其分布图，并分别求出其重心；



步骤二：我们作如下定义：平均任务竞争会员数 m_h= 会员预定总限额/任务总数，平均任务竞争会员数的含义是：每个任务平均有多少个会员去竞争。经计算可知，m_h=15，即每个任务有 15 名会员竞争。并且之前我们假设任务的定价与会员和任务的距离成线性关系，更加具体一点就是，会员和任务的距离越大，任务定价越高。当任务的定价无限增大，自然可以吸引所有的会员来竞争它，但是这显然是不合理的。所以我们需要让任务的定价尽可能小，但是又能够吸引一定的会员来竞争它。因此我们假设竞争每个任务的 15 个会员是最邻近该任务的 15 个人。具体操作如下：以每个任务为圆心，以不同的半径画圆，保证每个圆内的会员数为 15，将圈内的会员成为与其竞争的会员。



步骤三：确定影响因素，可能对任务定价产生影响的变量有：任务位置到任务重心的距离 $|r_i^j - \overline{r^j}|$，$|r_i^w - \overline{r^w}|$；竞争会员的位置到会员重心的距离 $|h_k^j - \overline{h^j}|$，$|h_k^w - \overline{h^w}|$；竞争会员到任务的距离 $((h_k^j - r_i^j)^2 + (h_k^w - r_i^w)^2)^{\frac{1}{2}}$ 以及竞争会员预定时间 t_k，预定限额 e_k 以及信誉值 x_k，假设所有变量与任务定价都是线性关系，即

$$v_i^l = a^j|r_i^j - \overline{r^j}| + a^w|r_i^w - \overline{r^w}| + \sum_{k\in O}(q^j|h_k^j - \overline{h^j}| + q^w|h_k^w - \overline{h^w}| + x_k(p_3((h_k^j - r_i^j)^2 + (h_k^w - r_i^w)^2)^{\frac{1}{2}} + p_1 t_k + p_2 e_k) + c$$

问题划归为求解系数 $a^j, a^w, q^j, q^w, p_1, p_2, p_3, c$

步骤四：假设给出的任务定价均为精准定价，也即不考虑任务完成情况。选理论定价与实际定价差的绝对值之和作为衡量标准，采用线性回归，求解出最优系数。

步骤五：对每个任务，定义增值比 $d_i = (v_i^s - v_i^l)/v_i^l$，根据拟合的结果，可以求出每个任务的增值比，之后用增值比来分析任务完成和失败的原因。

2）定价模型好坏的量度系数选取的是否吻合，我们采用绝对差算法来刻画，数学描述如下：

$$\sum_{k=1}^{N} |v_k^s - v_k^l|$$

N 表示任务总数

v_i^s 表示第 i 个任务的实际定价

v_i^l 表示第 i 个任务的理论定价

3）模型求解

（1）利用线性回归求出系数最优解。

以任务 i 为例，与其定价 v_i^l 有关的变量：$|r_i^j - \overline{r^j}|$，$|r_i^w - \overline{r^w}|$（任务的位置信息），$|h_k^j - \overline{h^j}|$，$|h_k^w - \overline{h^w}|$（竞争会员位置信息），$((h_k^j - r_i^j)^2 + (h_k^w - r_i^w)^2)^{\frac{1}{2}}$（竞争会员与任务的距离），$t_k$，$e_k$，$x_k$（竞争会员信息）.

假设所有的变量成线性关系，也即

$v_i^l = a^j|r_i^j - \overline{r^j}| + a^w|r_i^w - \overline{r^w}| + \sum_{k\in O}(q^j|h_k^j - \overline{h^j}| + q^w|h_k^w - \overline{h^w}| + x_k(p_3((h_k^j - r_i^j)^2 + (h_k^w - r_i^w)^2)^{\frac{1}{2}} + p_1 t_k + p_2 e_k) + c$

则我们的目标是选取系数 $a^j, a^w, q^j, q^w, p_1, p_2, p_3, c$，使得 $\sum_{k=1}^{n} |v_k^s - v_k^l|$ 达到最小。

用 MATLAB 求解结果如下：

$a^w = -1.258559895431065$

$a^j = 4.475039754284610$

$q^j = 0$

$q^w = 0$

$p_1 = -0.000847039018016$

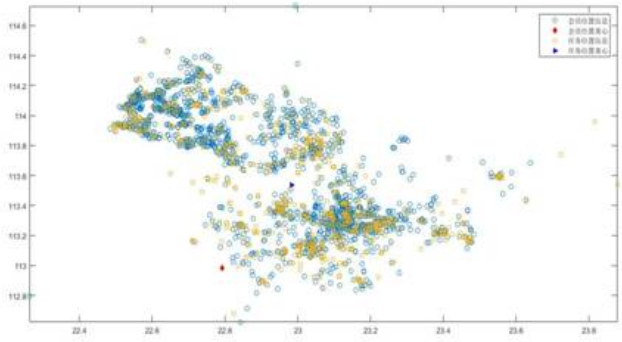
$p_2 = 0.000000529712632$

$p_3 = 0.006211717598381$

$c = 68.079526475834584$

下面根据解的情况给出相应解释。首先可以看出，a^w 为负数，即 $|r_i^w - \overline{r^w}|$ 越大，定价就越低；而 a^j 为正数，即 $|r_i^j - \overline{r^j}|$ 越大，定价就越高。而且 $|a^j| > |a^w|$，也就是说经度变化带来的影响要比纬度的大。根据深圳市地形地貌，我们给出如下解释。

深圳市地图



任务分布图

从地图上可以看出，任务位置的分布与经济繁华地带的分布基本一致。深圳市繁华地段主要沿经线，呈东西条带状分布。因此我们可以根据纬度的差异将繁华地带分成若干个区域，每个区域近似看成沿着经线分布的狭长地带，在每个区域内任务位置沿着经线分布。对于会员而言，他们偏爱选择邻近的任务，也即是说，他们更加偏好选取和他们在一个区域内的任务，而不愿意跨区域去执行任务，因此，任务的定价与纬度呈负相关，所以 a^w 的符号为负。而在同一个区域内，会员可能选择任何一个任务去执行。对于远离重心的任务，相对来说比较偏远，只有给出的酬金更高，才能吸引会员去执行它，因此任务的定价与经度成正相关，所以 a^j 为正数。并且在繁华地带，正所谓寸土寸金，经度的微小变化，都会带来周边经济实力的巨大差异，所以 $|a^j|$ 的数值较大。此外会员自身位置对于价格无影响，但是会员和任务的相对位置有影响，这和认知相符。并且从其他系数的正负可以看出，这组解与之前的假设：会员和任务的相对距离越远，会员的信誉度越高，预定限额越大，预定时间越短，会员执行任务所需的酬金

越高，相符，所以我们可以认为这组解是可行的。并且从系数的大小可知，影响任务定价的主要是位置因素。

（2）分析失败原因。

我们定义 $d_i = (v_i^s - v_i^l)/v_i^l$ 为任务 i 的增值比。由增值比的正负，我们可以看出任务的实际定价是高于还是低于理论定价；由增值比绝对值的大小，可以知道任务提价抑或降价幅度的大小。下面我们利用增值比将任务分为四种：

第 I 类表示 $d_i > 0$，且任务完成； 第 II 类表示 $d_i > 0$，任务失败；

第 III 类表示 $d_i < 0$，但任务完成；第 IV 类表示 $d_i < 0$，任务失败。

将任务分类完成后，我们分别计算每一类中任务增值比的平均值，分别记为 b_I，b_{II}，b_{III}，b_{IV}。MATLAB 计算结果如下：

第 I 类	第 II 类	第 III 类	第 IV 类
b_I=0.06546	b_{II}=0.04972	b_{III}=-0.03433	b_{IV}=-0.04393

由此可见 b_I>b_{II}，b_{III}>b_{IV}。所以对于失败的任务可做如下的解释：对于第 II 类的任务，虽然增值比大于零，即实际定价高于理论定价，但是由于它们价格的增幅还不够大，竞争力不够强，还不足以吸引会员来领取它，所以导致失败。而第 IV 类的任务，增值比小于零，即实际定价低于理论定价，并且相比于同样降价的任务，降价幅度过大，让会员无法接受，所以任务失败。（如表 1，表 2 所示）

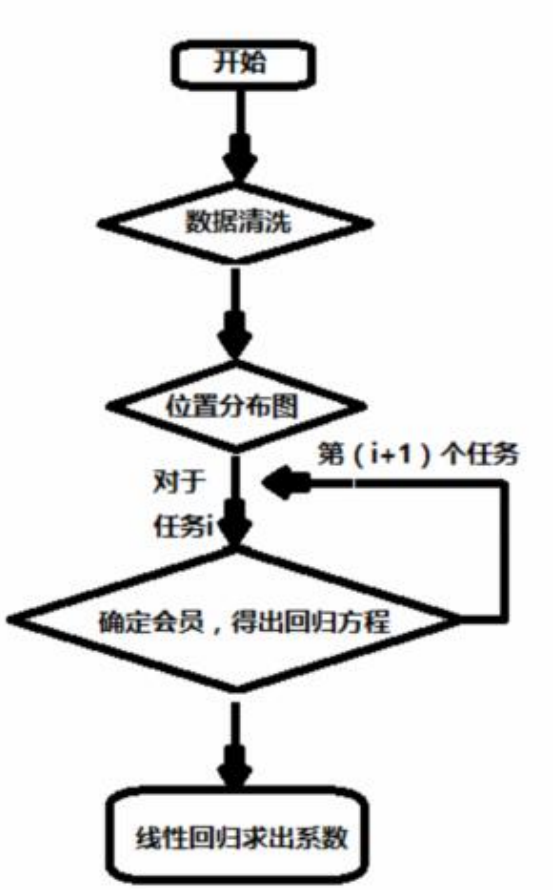
除了上述的失败原因分析，我们还能做更深入的思考：对于第 II 类的任务，他们失败的原因是价格增幅还不够大，所以我们可以把 b_I 作为一个阈值，当这些任务的价格增值比达到 b_I 时，即可认为第 II 类的任务由原来的失败变为了成功，同样的 b_{III} 可以作为第 IV 类任务的阈值。我们可基于这样的思考，对它们的价格进行调整，具体做法将在第二问详述。

任务号码	任务 gps 纬度	任务 gps 经度	任务标价	任务执行情况	增值比
A0398	23.37803295	113.2569259	75	1	0.068797
A0400	23.40329889	113.2398514	75	1	0.067734
A0225	23.33887144	113.1110649	75	1	0.067394
A0393	23.446661	113.2201755	75	1	0.065477
A0157	23.14872294	113.5152914	73.5	0	0.053621
A0169	23.38627895	113.410599	74	0	0.049487
A0446	22.68155741	113.9462038	72	0	0.048663
A0596	22.94081692	113.0626781	72.5	0	0.048107

$d_i > 0$ 时任务完成情况的对比

任务号码	任务 gps 纬度	任务 gps 经度	任务标价	任务执行情况	增值比
A0615	22.97994563	114.0020847	67	1	-0.033637
A0034	22.5829933	114.1471227	66	1	-0.034192
A0534	22.99313225	113.7323716	67	1	-0.034231
A0759	23.04235254	113.7318371	67	1	-0.034912
A0095	22.60766979	113.8685312	65.5	0	-0.044379
A0118	22.607611	113.8698943	65.5	0	-0.044489
A0608	23.01281035	113.1202326	66	0	-0.045097
A0391	23.13562043	113.3048107	65	0	-0.045216

$d_i < 0$ 时任务完成情况的对比



问题 1 流程图

5.2 问题 2 分析

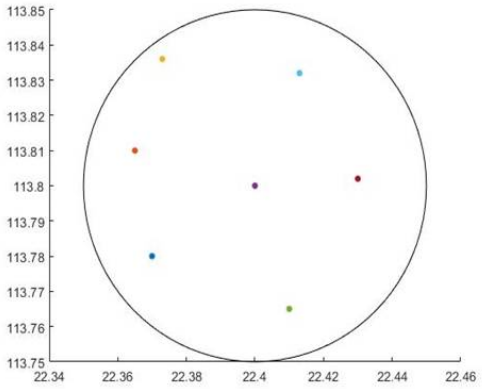
该问题要求我们对于附件一中的项目设计新的定价方案，并与原方案相比较。此时便会面临两个难题：一是从表格一，我们可以看出按照表格一中的定价，并不是所有的任务都完成了。这说明，附件一中所给的定价，并不能视为阈值定价，只能看成是阈值定价的一个估计限，因此我们需要找出一个阈值定价，当定价 $\geq$ 阈值，任务完成，$\leq$ 阈值，任务失败，以此来判断定价变化后，任务的完成情况；二是存在一些与认知相不符的情况，比如定价比理论值高，但是仍失败，定价比理论值低，却成功。所以这促使我们对任务进行聚类处理，考察局部性质。

针对存在的问题，便可以从多个角度来解答：第一个角度是在问题一当中，我们已经根据实际的定价拟合出了一个模型，此时能容易定义一个阈值定价，在此基础上对价格进行微调，提高任务完成率，这可视为一个新的定价方案；第二个角度是，对问题一中的模型进行改进。在问题一当中，我们将会员的预选限额，预选时间以及信誉度当成独立的变量，而在题目当中提到，信誉度越高，预选时间越长，预选限额也越大，所以我们可以考虑这三个量之间的相互关系，以此来优化模型，重新定价；第三个角度是不改变问题一中的模型，利用线性规划的知识重新选定系数，使其达到最优。

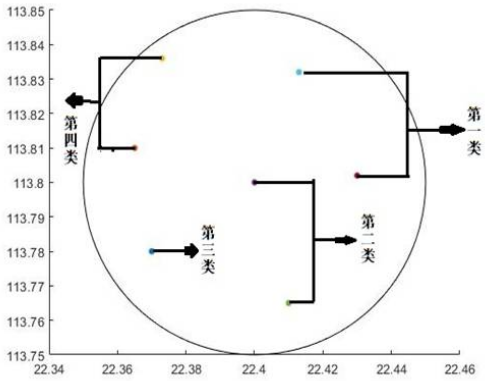
具体的分析如下。

角度一：根据问题一的结果微调价格由问题一最后的分析可以发现，第 II 类，第 IV 类的任务是失败的任务，根据之前的解释，如果我们分别将 b_I，b_{III} 当成第 II 类，第 IV 类的任务的阈值，则我们希望能够对这些失败任务的价格进行微调，使之达到阈值。相反的，对于第 I 类和第 III 类这些成功的任务，有些增值比大于 b_I，b_{III}，说明这份多出的价格是多余的，所以我们希望在确保这些任务仍然能被完成的情况下，适当降低价格。但是如果我们贸然直接使用 b_I 和 b_{III} 去给任务重新定价，则会出现任务 i 属于第 I 类的任务，但是 $d_i < 0$，也即此时无法从这个成功完成的任务中拨出多余的价格。所以为了更加精细的分析，我们对于每个任务计算其局部阈值。

1）局部范围大小的选取我们定义：平均竞争任务数 m_r = 会员预定总限额/会员总数，平均竞争任务数的含义是多少个任务去竞争一个会员领取它。经 matlab 计算 $m_r = 7$，即 7 个任务去竞争一个会员领取它。从问题一中的回归模型可知，影响任务定价主要是位置因素，所以每个任务主要是和邻近的任务去竞争会员。所以我们在地图上以每个任务画圈，使得圈内的任务数为 7，圈内的任务称为竞争任务。



2）局部阈值的确定以任务 i 为例，对于圆圈 i 内的 7 个竞争任务，将其归为四类，即之前所说的第 I 类，第 II 类，第 III 类，第 IV 类。然后计算出这四类中的任务的增值比的平均值，将其作为任务 i 的局部阈值，记为 b_{Ii}，b_{IIi}，b_{IIIi}，b_{IVi}。



3）价格的调整根据以上的叙述，下面给出，价格微调的标准：1. 对于失败的任务，如果 $d_i > 0$，则将其与 b_{Ii} 相比较，如果 $< b_{Ii}$, 则提升价格使之达到阈值，如果 $> b_{Ii}$, 则不作调整；如果 $d_i < 0$；则将其与 b_{IIIi} 相比较，如果 $< b_{IIIi}$，则提升价格使之达到阈值，如果 $> b_{IIIi}$，则不作调整。2. 对于成功的任务，如果 $d_i > 0$，则将其与 b_{Ii} 相比较，如果 $< b_{Ii}$, 则不作调整，如果 $> b_{Ii}$, 降低价格使之达到阈值；如果 $d_i < 0$，则将其与 b_{IIIi} 相比较，如果 $< b_{IIIi}$，则不作调整，如果 $> b_{IIIi}$，降低价格使之达到阈值。例如：

任务号码	任务 gps 纬度	任务 gps 经度	任务标价	任务执行情况	调整之后的价格
A0001	22.56614225	113.9808368	66	0	68.362
A0002	22.68620526	113.940525 2	65.5	0	69.424
A0003	22.57651183	113.957198	65.5	1	64.809
A0004	22.56484081	114.2445711	75	0	75
A0005	22.55888775	113.9507227	65.5	0	67.771
A0006	22.55899906	114.2413174	75	0	75
A0007	22.54900371	113.9722597	65.5	1	65.5
A0008	22.56277351	113.9565735	65.5	0	66.716
A0009	22.50001192	113.8956606	66	0	69.065
A0010	22.5437861	113.9239778	66	1	65.711
A0011	22.52486369	113.9308596	65.5	0	67.713
A0012	22.519087	113.9358436	65.5	0	68.416
A0013	22.54797243	113.977909	65.5	1	65.5

图中价格提升的标记为蓝色，价格下降的标记为红色，价格不作调整的标记为绿色。

从表格中还能获取以下的信息：

1. 重新定价的任务数占总任务数的 60%，也即绝大多数任务将被重新定价，由此可知模型对绝大多数任务有效，可见模型的合理性；

2. 在价格未作调整的任务中，原先未完成的任务只有 65 个，占总任务数的 8%。这些任务可视为价格调整过后无法被完成的任务。所以可知，价格调整过后，任务完成率将达到 92%，由此可见模型的优越性。

3. 经过统计可知，提升价格所需的资金为 652 元，而多出的价格只有 289，之间的落差是 APP 平台需要补入的资金。

优缺点分析：

优点：能够给出一个阈值标准，定量精准的判断价格变化之后任务的执行情况；

缺点：只是在原有的基础上对价格进行微小的调整，没能从根本上提出更为优越的定价方案。

角度二：优化问题一中的线性模型对问题一，我们采用的线性模型中，将会员的预选限额，预选时间以及信誉度当成独立的变量，但是在题目当中指出，信誉度越高，预选时间越长，预选限额也越大，所以我们在优化问题一中的模型时，可以考虑这三个量之间的相互关系，重新定价，具体做法如下

1）模型的优化

原先的模型是

$v_i^l = a^j|r_i^j - \overline{r^j}| + a^w|r_i^w - \overline{r^w}| + \sum_{k\in O}(q^j|h_k^j - \overline{h^j}| + q^w|h_k^w - \overline{h^w}| + x_k(p_3((h_k^j - r_i^j)^2 + (h_k^w - r_i^w)^2)^{\frac{1}{2}} + p_1t_k + p_2e_k) + c$

现在加入 x_i 对 e_i 和 t_i 的影响之后，模型变为

$v_i^l = a^j|r_i^j - \overline{r^j}| + a^w|r_i^w - \overline{r^w}| + \sum_{k\in O}(q^j|h_k^j - \overline{h^j}| + q^w|h_k^w - \overline{h^w}| + x_k(p_3((h_k^j - r_i^j)^2 + (h_k^w - r_i^w)^2)^{\frac{1}{2}} + p_1(t_k + n_1x_k) + p_2(e_k + n_2x_k)) + c$

因此我们的目标变为是找到一组系数 a^j,a^w,q^j,q^w,p_3,p_1,p_2 ,n_1,n_2,c，使得 $\sum_{k=1}^{n}|v_k^s - v_k^l|$ 达到最小。

2）重新拟合

和问题一中的做法一样，使用线性回归，但是会发现，现有的方程不足以解出 n_1,n_2, 所以，我们保证 $n_1^2 + n_2^2$ 达到最小, 最后求解得到

$$n_1 = 0 \qquad n_2 = 0.00335$$

这说明，x_i 对 e_i 有正向的影响，而对于 t_i 无影响。和题目中所说的基本吻合。

并且重新拟合之后，会发现新的理论价格和原始的理论价格有不同。

任务号码	任务 gps 纬度	任务 gps 经度	任务标价	任务执行情况	原始理论价格	新的理论价格
A0001	22.56614225	113.9808368	66	0	67.139	66.862
A0002	22.68620526	113.9405252	65.5	0	68.678	68.864
A0003	22.57651183	113.957198	65.5	1	65.117	64.791
A0004	22.56484081	114.2445711	75	0	68.629	68.779
A0005	22.55888775	113.9507227	65.5	0	67.12	65.497
A0006	22.55899906	114.2413174	75	0	68.68	68.787
A0007	22.54900371	113.9722597	65.5	1	67.342	66.07
A0008	22.56277351	113.9565735	65.5	0	65.845	64.222
A0009	22.50001192	113.8956606	66	0	68.314	68.465
A0010	22.5437861	113.9239778	66	1	68.091	67.934
A0011	22.52486369	113.9308596	65.5	0	68.24	68.321
A0012	22.519087	113.9358436	65.5	0	68.204	68.249
A0013	22.54797243	113.977909	65.5	1	67.621	66.185
A0014	22.50616871	113.9314284	66	1	68.216	68.268
A0015	22.49962566	113.9365145	66	1	68.221	68.271
A0016	22.54032142	113.9236456	66	1	68.167	68.149

3）根据新的模型重新计算增值比

根据新的理论价格，我们可以重新计算增值比，然后沿用上述的理论，重新计算阈值，然后对价格进行微调。这里不再赘述，只附上新的增值比。

任务号码	任务 gps 纬度	任务 gps 经度	任务标价	任务执行情况	原始增值比	新的增值比
A0001	22.56614225	113.9808368	66	0	-0.016962	-0.012899
A0002	22.68620526	113.9405252	65.5	0	-0.046274	-0.048845
A0003	22.57651183	113.957198	65.5	1	0.0058834	0.010948
A0004	22.56484081	114.2445711	75	0	0.092828	0.090443
A0005	22.55888775	113.9507227	65.5	0	-0.024142	0.0000513
A0006	22.55899906	114.2413174	75	0	0.092014	0.090328
A0007	22.54900371	113.9722597	65.5	1	-0.027352	-0.008623
A0008	22.56277351	113.9565735	65.5	0	-0.0052464	0.019906
A0009	22.50001192	113.8956606	66	0	-0.033875	-0.036005
A0010	22.5437861	113.9239778	66	1	-0.030706	-0.028466
A0011	22.52486369	113.9308596	65.5	0	-0.040156	-0.041295
A0012	22.519087	113.9358436	65.5	0	-0.039639	-0.040285
A0013	22.54797243	113.977909	65.5	1	-0.03136	-0.010349
A0014	22.50616871	113.9314284	66	1	-0.03249	-0.033222
A0015	22.49962566	113.9365145	66	1	-0.032557	-0.033267
A0016	22.54032142	113.9236456	66	1	-0.031787	-0.031538

优缺点分析：

优点：优化了原有的模型，沿用了上述的方法。

缺点：只是在原有的基础上对价格进行微小的调整，没能从根本上提出更为优越的定价方案。

角度三：线性规划选择最优系数

根据上述的讨论可知，我们假设定价模型是一个线性模型，最终的定价取决于系数的选择。并且由表一的数据我们知道，在实际的定价下，并不是所有的任务都完成，所以实际给出的价格不能作为阈值价格，只能作为阈值价格一个限界。所以我们做出如下假设：对于任务 i, 当任务完成时，认为理论价格 $v_i^l <$ 实际价格 v_i^s；当任务失败时，认为理论价格 $v_i^l >$ 实际价格 v_i^s. 例如：

对第 833 个任务

任务号码	任务 GPS 纬度	任务 GPS 经度	任务标价	任务执行情况	原始增值比
A0833	22.81467597	113.8277312	85	1	0.23214

有

$$v_{833}^l = a^j|r_{833}^j - \overline{r^j}| + a^w|r_{833}^w - \overline{r^w}| + \sum_{k\in O}(q^j|h_k^j - \overline{h^j}| + q^w|h_k^w - \overline{h^w}| + x_k(p_3((h_k^j - r_{833}^j)^2 + (h_k^w - r_{833}^w)^2)^{\frac{1}{2}} + p_1t_k + p_2e_k) + c \leq 85$$

对第 101 个任务

任务号码	任务 GPS 纬度	任务 GPS 经度	任务标价	任务执行情况	原始增值比
A0101	22.73006657	114.0666477	66.5	0	-0.034279

有

$v^l_{101} = a^j|r^j_{101} - \overline{r^j}| + a^w|r^w_{101} - \overline{r^w}| + \sum_{k\in O}(q^j|h^j_k - \overline{h^j}| + q^w|h^w_k - \overline{h^w}| + x_k(p_3((h^j_k - r^j_{101})^2 + (h^w_k - r^w_{101})^2)^{\frac{1}{2}} + p_1t_k + p_2e_k) + c \geq 66$

这样可以得到 800 多个不等式，可以使用线性规划 [3] 中的单纯形法，求解出最优解。但是此时一个比较关键的问题是，我们无法找到一个阈值，从而很难判断在系数最优解的情况下，这些任务完成与否。所以该方法仅做参考。

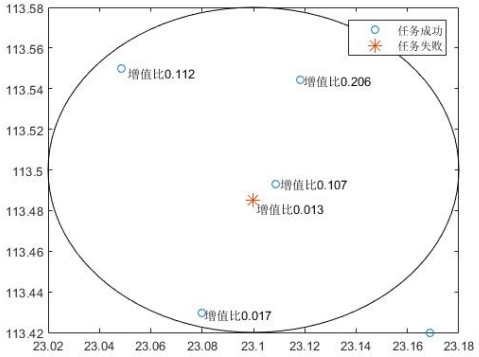
优缺点分析：

优点：另辟蹊径，从根本上优化了定价模型

缺点：算法运行效率低，难以求解，并且缺少进一步的理论确定出新的阈值来判断任务执行情况。

5.3 问题 3 分析

问题 3 中提到，由于一些任务位置较为集中，会员争相竞争，从而导致一些任务无人问津。这个现象可以从问题 2 中的优化角度一，利用增值比定量地看出。以任务 i 为例，任务 i 的实际价格比理论价格要高，即 $d_i > 0$，但是相比于周边的任务，增值比 d_i 较小，所以，任务 i 无人问津，任务失败。



问题出现的根本原因是，在任务密集的区域，由于竞争激烈，增值比上的微小差异被放大，最终给任务的完成与否带来的决定性的影响。想要解决这个矛盾，很自然的想法是把位置邻近，增值比相近的任务打包处理，用整体的增值比代替单个的增值比，这样能消除任务之间增值比上的微小差异，从而避免竞争带来的不利影响。

但是在具体提出打包分类方案时，我们会面临以下的难题：

1. 每个任务包中包含多少任务最宜；

2. 对任务进行聚合分类时，优先顺序该如何确定；

3. 具体的分类标准是什么。

下面我们就逐步回答上述问题，提出我们的打包分类方案：

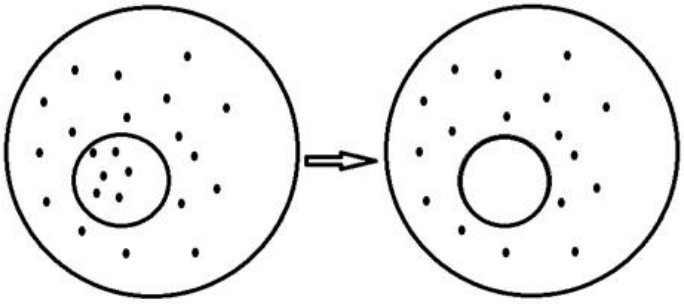
Step 1：每个任务包中包含多少任务最宜

每个任务包中任务数的多少对于分类模型最终的效果有着决定性的的影响：如果任务包中任务数过多，则由于会员预定限额的限制，选取这个任务包的会员就会很少，因此对任务包而言，选择执行会员的空间减少。并且任务过多，毫无疑问会降低任务完成的效率；而如果任务数过少，则体现不出打包分派任务的优越性。因此我们必须合理选择任务包中的任务数。通过分析附件二中的数据，从会员的预定限额可以看出，如果将所有的任务都打包处理，则对于一些任务预定限额很少的会员，将没有机会领取任务，这是人力资源的浪费，所以我们没有必要对所有任务都打包处理。通过上面的分析，可以综述如下：对于任务包中的任务数，我们需要找到一个上限。通过权衡，我们决定选取平均竞争任务数 m_r= 会员预定总限额/会员总数，作为包裹中任务数的上限。m_r 在前文已经使用过，但是在这里 m_r 的含义略有不同，解释为平均每个会员能够领取到多少个任务，由上文的计算可知，$m_r = 7$.

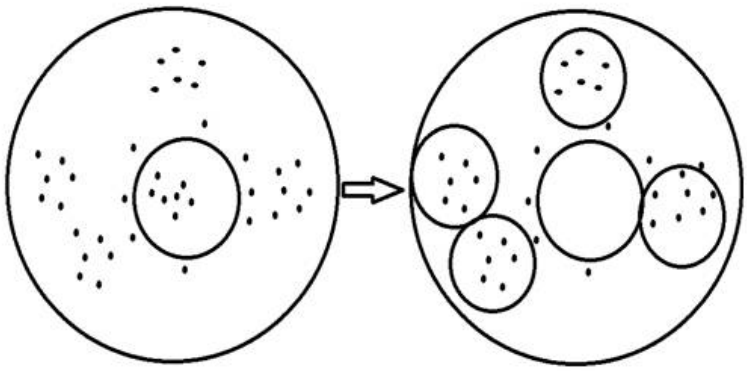
Step 2: 打包优先顺序的决定

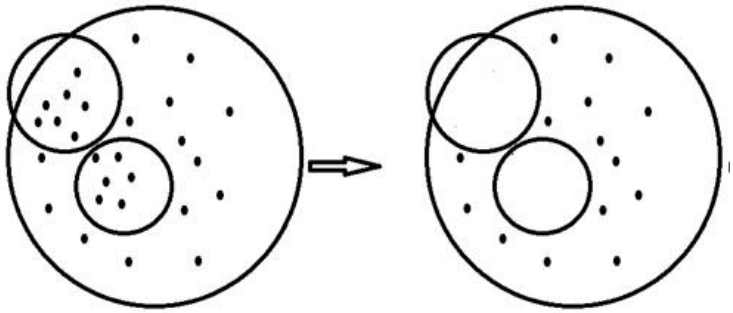
任务打包的优先顺序与任务的位置分布息息相关。再次观察任务分布图，结合拓扑学和离散数学的知识 [4]，我们可以将任务分布形成的集合近似成一个单连通的区域。即此时区域是连通的，且内部没有空洞，边界只有外围的一圈。

如果我们从内部开始分类打包，则会出现如下的情形：每次打包完一个任务包，就等于在原来的区域内部挖去一块，留下一个空洞，并且新增了一圈边界。如图：



空洞的出现破坏了原来区域的单连通性，当空洞数目过多，原来的区域则被分割成互不相交的几块，这种情形不仅给算法的实现带来了巨大的困难，而且对于在分割出的区域内的任务，可能无法再使用之前的打包分类原则进行打包。如图：





此外，对于新增边界处的任务，对它们进行打包也有不小的困难，如图：

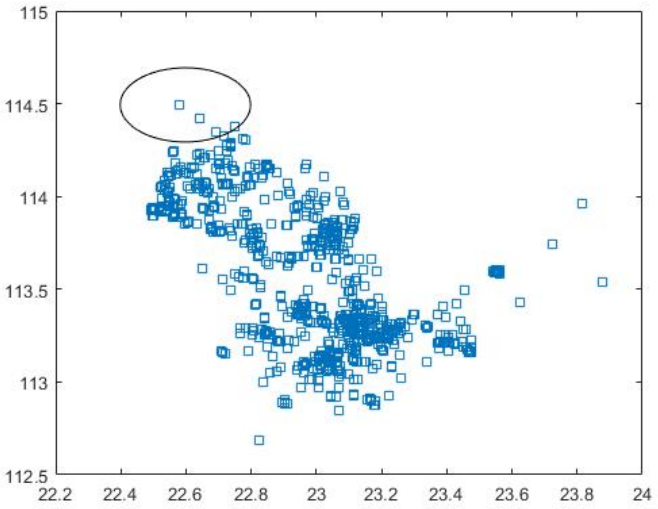
基于以上的考虑，我们决定采用从外围开始打包的策略 [5]。

Step 3：具体打包方案

重心转移法

算法思路：

根据 step2，我们采取从外围开始打包的策略。更具体地讲，我们选择从离重心最远的那个点着手。因此，首先求出整个任务分布位置的重心，然后计算距离，求出距离重心最远的那个任务。接下来以这个任务为中心，在这里我们设置了一个最大距离 l，以 l 为半径画圆，如果圆圈内的任务数不超过 7，则将圈内的任务打包；如果圆圈内的任务超过 7，则减小半径，使得圈内的任务数小于 7 为止。选取完毕后，将圈内的任务视为一个包裹。



这时从整体的任务中除去这些已经打包的任务，得到一个新的集合，重新计算这个新的点集的重心位置，重复上述过程，便能将所有的任务进行分类打包.

打包结果, 例如：

包的编号	包内的任务编号	包的重心	包的实际价格
1	A0450 A0038 A0063 A0050	22.638 114.421	305
80	A0157 A0210 A0149 A0148 A0150 A0182 A0129	23.138 113.458	477.5
100	A0688 A0735 A0550 A0167 A0788 A0176 A0553	23.217 113.185	507.5

进一步的拟合：

将任务打包完成之后，我们建立新的线性回归模型来探索对任务完成的情况

（1）模型的建立

以第 i 个任务包为例，在这个任务包中共有 k 个任务（$k \leq 7$），我们利用数据聚合作如下的定义：

$$V_i^s = \sum_{m \in O} v_m^s$$

$$R_i^j = \frac{\sum_{m \in O} r_m^j}{k}$$

$$R_i^w = \frac{\sum_{m \in O} r_m^w}{k}$$

例如：

包的编号	包内的任务编号	包的重心	包的实际价格	包的理论价格
1	A0450 A0038 A0063 A0050	22.638 114.421	305	285.114
80	A0157 A0210 A0149 A0148 A0150 A0182 A0129	23.138 113.458	477.5	474.413
100	A0688 A0735 A0550 A0167 A0788 A0176 A0553	23.217 113.185	507.5	487.26

和问题一一样，假设 V_i^s 与变量 $|R_i^j-\overline{r^j}|$，$|R_i^w-\overline{r^w}|$，$((h_k^j-r_i^j)^2+(h_k^w-r_i^w)^2)^{\frac{1}{2}}$，$t_m$，$e_m$ 成线性关系, 模型的最终目的是确定出最优的一组系数 A^j，A^w，P_1，P_2，P_3，C

（2）数学描述

$V_i^l=A^j|R_i^j-\overline{r^j}|+A^w|R_i^w-\overline{r^w}|+\sum\limits_{k\in O}x_m(P_3((h_m^j-R_i^j)^2+(h_m^w-R_i^w)^2)^{\frac{1}{2}}+P_1t_m+P_2e_m)+C$

使得 $\sum\limits_{k=1}^{n}|V_k^s-V_k^l|$ 求和达到最小

（3）模型的求解

利用线性回归求的系数

$$A^w=-21.1547696388467$$
$$A^j=44.0379709381243$$
$$P_1=-0.0815823312844$$
$$P_2=0.0005053407382$$
$$P_3=-0.0001669136836$$
$$C=827.2566024862618$$

下面对于所求的系数作出相应的解释：

与第一问中的系数相比，可以发现打包过后所有系数的值都得到增大，这和将整个包当成一个单位处理，实际价格和理论价格均增大的情况是相符的。通过进一步的研究发现 A^w 和 a^w，A^j 和 a^j 符号仍然相同，且相对大小基本不变，所以在打包过后不改变问题一所得出的定价规律。对于 P_1 和 p_1，P_2 和 p_2，他们符号相同，且相对大小也基本不变，只有 P_3 符号改变但是 P_3 值相对很小，对结果的影响可以忽略不计，因此可以认为这组系数符合要求。

（4）对任务完成情况的影响

在上述基础上，我们对这个模型的数据作出如下分析。

打包处理过后，我们重新计算每个任务包的增值比，并用任务包的增值比来作为一个任务包能否被完成的评判依据。例如：

包的编号	包内的任务编号	包的重心	包的实际价格	包的理论价格	包的增值比
1	A0450 A0038 A0063 A0050	22.638 114.421	305	285.114	0.07
80	A0157 A0210 A0149 A0148 A0150 A0182 A0129	23.138 113.458	477.5	474.413	0.007
100	A0688 A0735 A0550 A0167 A0788 A0176 A0553	23.217 113.185	507.5	487.26	0.042

根据之前的分析，我们采用任务包的增值比来代替单个的增值比，是希望消除任务之间增值比上的微小差异，从而避免竞争带来的不利影响。但是注意到，任务包中的任务原先的完成

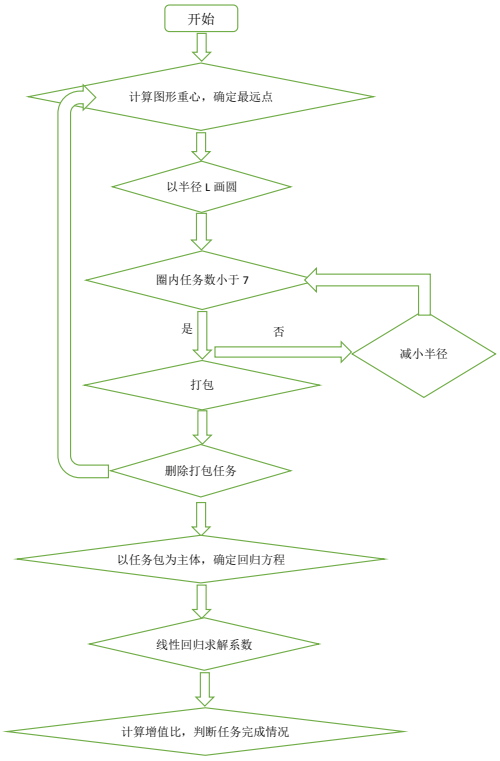
情况可能是参差不齐的，而打包处理后，他们的完成情况将被绑定，要么同时被完成，要么同时失败。所以这时选取一个合适的阈值来判断一个任务包完成与否显得尤为重要。

经过尝试可知，如果我们仍然选取问题 2 中的阈值 b_I，b_{III} 作为评判标准，将任务包的增值比 D_i 直接与之比较，以此来判断任务完成情况，则会得到结果：46 个任务包成功，57 个任务包失败。这个结果显然不是我们想要的。并且通过分析就能发现上述的判断方法有很大的弊端：因为当任务聚类成任务包后，虽然无法完成的任务的不利因素得到了稀释，但同时整个任务包的性能必定比包内原先就能被完成的任务的性能要低，如果仍然采用之前评判单个任务的评判标准，显然是不合理的。

我们采用如下的判断方法：我们计算每个任务包中任务增值比的平均值，将这个平均值与任务包的增值比相比较。如果任务包增值比 $D_i >$ 这个平均值，则意味着在打包过后，总体的性能得到了提高，此时任务包被完成的可能性变大，因此我们将这个任务包视为成功。计算可得，打包过后，共有 73 个任务包被完成，51 个任务包失败，

（5）结论

经过打包处理，消除了增值比在邻近任务之间的微小差异，避免了恶性竞争的出现。从算法结果来看打包过后任务完成率总体略有提高。但是，此模型仍然具有缺陷，最为关键的是缺少进一步的理论来确定阈值的选取，更为精确的判断任务的执行情况。



问题 3 流程图

5.4 问题 4 分析：

从附件三我们可以看到，此次我们需要为两千多个任务定价。由前三问可知，我们可以选用问题一中的线性模型，为它们挨个定价，也可以采用问题三中的策略，先打包再对任务包定价。而此处，我们决定使用问题三中的策略，理由如下：

1. 此次任务数目是原先的两倍多，竞争更为激烈，使用打包处理的策略能够避免恶性竞争，提高任务完成率。

2. 此次会员需要执行的平均任务数增多，使用打包策略，能减少任务分配的时间，提高效率。

3. 随着任务数的增多，问题一中的线性模型拟合匹配度降低，对于问题 4 可能不再适用。

因此，我们延继着前述的打包模型，先对任务打包，再利用前述的线性模型和求出的系数数据，可以得到每个包的定价。

例如

打包结果：

包的编号	包的任务
1	C1166 C1146 C1014 C0011 C0034 C0039 C0046
150	C1966 C0891 C1964 C1965 C1956 C1962 C0971
300	C0629

定价方案：

包的编号	包的任务	包的价格
1	C1166 C1146 C1014 C0011 C0034 C0039 C0046	495.782
150	C1966 C0891 C1964 C1965 C1956 C1962 C0971	482.954
300	C0629	87.732

6 模型评估和改进

6.1 模型的优点

（1）选用线性模型探索定价规律，符合线性定价原则，基本探索出定价规律，为问题的后续深入奠定基础。

（2）采用外围收缩，重心转换的方法进行任务聚类，大大减小了计算复杂度，可操作性强。

6.2 模型的缺点

（1）未能充分利用数据信息，比如未能使用线性规划模型求解出最优的系数解，仍需对所给数据进一步挖掘。

（2）需要进一步改进任务聚类的方法, 从外围收缩的算法是建立在将所有任务都打包的基础之上，在一些情况下，对于极端的边界点，可以选择剔除，在现有的模型中并没有给出这样的评判标准。

6.3 模型的改进和推广

（1）改进：在已有的打包模型的基础上，可以只打包一部分任务点，对于剩下的任务点，用 k 近邻算法判断其所属的包。这避免了在判断较靠内部的点时，只由距离单个点的相对距离决定打包的缺陷。

（2）推广：在模型中添加非线性因素进行优化，使得模型可以考虑更多的因素。这样可以参考引用更多的数据，使得模型更加贴近现实。

参考文献

[1] 罗荣华, 线性定价法则, 西南财经大学金融学院.

[2] 姜启源，谢金星，叶俊. 数学模型 [M]. 北京：高等教育出版社，2003.

[3] 张干宗. 线性规划 [M]. 武汉大学出版社，2010.

[4] 尤承业. 基础拓扑学讲义 [M]. 北京大学出版社，1997.

[5] 薛申芳.MATLAB 在平面图形边界信息提取中的应用 [J]. 衡水学院学报.2009.