“拍照赚钱”的任务定价

摘要

本文针对“拍照赚钱”的任务定价问题，结合基于遗传模拟退火算法的聚类分析法和偏最小二乘法，研究出任务定价规律并分析得到任务未完成原因；建立了基于任务难度的定价函数模型、ELM 极限学习机模型和定价综合评价模型，设计了新的任务定价方案，并和原方案进行比较；建立了基于任务难度、颗粒度和预定限额方差指标的定价函数模型、ELM 极限学习机模型和定价综合评价模型，解决了任务打包情况下的定价、评价、预测的问题；建立了新增地区经济指标的定价函数模型，对新项目定价并评价。

针对问题一，结合基于遗传模拟退火算法的聚类分析法和偏最小二乘法，研究出任务定价规律并分析得到任务未完成原因。首先进行数据可视化处理，找出影响价格的因素；其次使用基于遗传模拟退火算法的聚类分析对区域进行划分，定量化指标值；最后使用偏最小二乘法分别针对价格和任务完成率根据各指标构建线性回归模型，并且通过相关性分析得出任务未完成的主要原因。通过计算残差平方和对模型检验，发现拟合效果较好。结果显示，任务价格与任务密度、会员密度存在线性回归关系，影响任务完成情况的主要因素是会员密度。

针对问题二，建立了基于任务难度的定价函数模型、ELM 极限学习机模型和定价综合评价模型，设计了新的任务定价方案，并和原方案进行比较。首先提出任务难度指标并在转换思路的方式下简明地定量了指标；其次按照机理性分析任务难度指标与定价的关系并考虑函数特征，构建了类似于 Sigmoid 型的定价函数；再次构建极限学习机模型，按照经纬度和价格，对任务完成情况进行仿真预测；最后根据新定价方案的仿真任务完成率、平均价格从效益成本两方面构建综合评价模型分析两种定价方案的优劣。通过分析极限学习机分类正确率对模型进行检验，求得正确率较高，模型较为适宜。最终得出定价函数模型，并且求得新老方案的综合指标分别为 0.54、0.46，分析可知新定价方案的效果更好。

针对问题三，建立了基于任务难度、颗粒度和预定限额方差指标的定价函数模型、ELM 极限学习机模型和定价综合评价模型，解决了任务打包情况下的定价并评价的问题。首先考虑到此问题新增的任务打包发布情况，可以引入任务颗粒度指标，由于各个会员预定限额的差距会引起类似于垄断的效应，还应引入预定限额方差指标，构建定价函数；其次对其使用极限学习机模型进行仿真预测；最后通过综合评价指标分析任务打包发布的优劣性。进行灵敏度分析，发现任务难度对价格最灵敏。最终修改得出新的定价函数模型，此时任务完成率增加到 82%，由于打包发布与否的综合指标分别为 0.50、0.47，分析可知任务打包发布效果更好。

针对问题四，建立了地区因素定价模型，解决了新项目的任务定价并评价的问题。对于新项目，分析任务分布的实际情况，并且考虑打包发布任务，在上述定价模型的基础上引入地区经济指标，修正定价函数。通过仿真预测以及综合评价就可以对新项目的定价方案进行评价。进行灵敏度分析，发现任务难度对价格最灵敏。最终求得新方案各任务的具体定价，在这种情况下定价方案综合指标与普通定价方案的综合指标分别为 0.54、0.46，分析可知此定价方案更优。

关键词：遗传模拟退火算法 偏最小二乘法 相关性分析 极限学习机 综合评价

一、问题重述

“拍照赚钱”是移动互联网下的一种自助式服务模式。用户下载 APP，注册成为 APP 的会员，然后从 APP 上领取需要拍照的任务（比如上超市去检查某种商品的上架情况），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。如果定价不合理，有的任务就会无人问津，而导致商品检查的失败。

附件一是一个已结束项目的任务数据，包含了每个任务的位置、定价和完成情况（“1”表示完成，“0”表示未完成）；附件二是会员信息数据，包含了会员的位置、信誉值、参考其信誉给出的任务开始预订时间和预订限额，原则上会员信誉越高，越优先开始挑选任务，其配额也就越大（任务分配时实际上是根据预订限额所占比例进行配发）；附件三是一个新的检查项目任务数据，只有任务的位置信息。请完成下面的问题：

1. 研究附件一中项目的任务定价规律，分析任务未完成的原因。
2. 为附件一中的项目设计新的任务定价方案，并和原方案进行比较。
3. 实际情况下，多个任务可能因为位置比较集中，导致用户会争相选择，一种考虑是将这些任务联合在一起打包发布。在这种考虑下，如何修改前面的定价模型，对最终的任务完成情况又有什么影响？
4. 对附件三中的新项目给出你的任务定价方案，并评价该方案的实施效果。

附件一：已结束项目任务数据
附件二：会员信息数据
附件三：新项目任务数据

二、问题分析

2.1 问题一的分析

针对问题一，要求研究附件一中项目的任务定价规律并分析任务未完成的原因。问题一是后续问题的基础，在分析价格影响因素及机制、任务未完成原因后，才能制定优秀的定价方案。问题一属于多元回归、相关性分析问题，通常使用最小二乘法、偏最小二乘法求解回归系数。在此考虑到因素间可能存在多重相关性，于是决定使用偏最小二乘法拟合数据，实际效果比传统的最小二乘法的结果好。

值得注意的是，由于价格和完成情况可能都与会员和任务的位置分布有关，而并非仅与任务的经纬度有关。为了能够简明、合理定义位置关系并定量化指标，引入任务密度和会员密度来定义分布关系。考虑到定量化密度，那么应当划分区域，在此采用聚类分析法划分区域，为了得到最好的聚类效果，此处采用基于遗传模拟退火算法的聚类分析。这样就能够简明地求出具体指标值，同时还可以求出各个地区任务完成率，从而可以拟合数据并进行相关性分析。

通过分析残差平方和以及相关性系数，对多元线性回归模型进行检验。该模型预期可以求出价格规律，以及影响未任务的最主要原因。经过这一系列分析，此模型是适宜的。

2.2 问题二的分析

针对问题二，要求为附件一中的项目设计新的任务定价方案，并和原方案进行比

较。该问题在这道题中有着承上启下的作用，可以参考问题一的定价规律以及众包平台定价的相关因素，构造合理的定价模型，同时为后续问题进行铺垫。问题二属于构造定价函数模型、仿真分类预测模型和综合评价模型的三模型综合问题。

其中仿真分类预测模型通常可以建立 SVM 支持向量机、LVQ 神经网络、ELM 极限学习机模型等，由于 ELM 极限学习机在分类预测中相比其余两个模型，有更快的运行速度和更高的正确率等,所以本文采用 ELM 极限学习机作为仿真分类预测模型。综合评价模型通常可以由主观赋权法、层次分析法和熵权法等方法建立，由于预期通过成本效益两方面提出指标，指标数量较少，仅结合众包平台机制即可主观赋权建立综合评价模型。

首先根据题目条件，引入合适的定价指标并在保证简明性的条件下，将指标定量化；其次按照机理性分析指标与定价的关系并考虑函数特征，构建定价函数模型；再次构建 ELM 极限学习机模型，按照经纬度和价格，对任务是否完成进行仿真分类预测；最后根据新定价方案的仿真任务完成率和新定价方案的平均价格，从效益成本两方面构建综合评价模型，分析两种定价方案的优劣。

此问题中 ELM 极限学习机模型十分重要，它不仅可以用来优化调节定价模型中的部分参数，还可以求解综合评价模型的任务完成率指标。通过 ELM 极限学习机进行机器学习，挖掘出隐藏在数据背后的灰色关系，良好地反映位置、价格对任务完成情况的影响。在此问题中不仅只考虑如何定价，更重要的是，无论定价函数是怎样的形式都能够较好地判别各种定价方案的优劣，有着良好的适应性。

根据 ELM 极限学习机的分类正确率以及综合评价指标的灵敏度分析，对此模型进行检验分析。预期可以得到新的定价函数，并且此定价方案效果较好。经过一系列分析，可知此模型是适宜的。

2.3 问题三的分析

针对问题三，要求确定考虑将多个任务打包后的定价方案，并评价实施效果。该问题与问题二相类似但又有不同，是问题二的升华，因为使用了任务打包发布，还应该建立相关的指标来修改定价函数。

首先分析打包问题带来的定价影响，推测定价函数与任务密度、任务颗粒度和预定限额方差有较大关系，所以在问题二的定价函数的基础上，引入部分新指标。通过控制变量并结合机理分析发现各指标对定价的影响，加入合适的函数项，对问题二中的定价函数进行修改。

其次建立基于极限学习机的仿真预测模型，对任务点的位置和原定价进行仿真测试，根据新的任务价格预测出新的规则下任务完成情况；最后建立综合评价模型，求解综合指标。

此定价模型估计会有较多评价指标，因此，需要进行灵敏度分析，找出影响最灵敏的指标。预期可以得到修改的定价函数，打包定价方案比不打包的效果好。经过一系列分析，可知此模型是适宜的。

2.4 问题四的分析

针对问题四，要求针对新项目给出任务定价方案，并评价该方案的实施效果。问题四是上述所有讨论的结合，是在探讨定价模型后的实际应用。

由于此问题是全新的应用，需要结合实际考虑。首先应做任务分布图，找出实际分布规律；其次根据图像考虑是否还存在影响定价的实际因素；再次结合实际因素对定价模型进行修正；最后对任务完成情况进行仿真分类预测，分析定价优劣。

推测可能需要结合地理位置，划分不同市区，根据地区经济状况等因素来对模型进行最终修正，再根据综合评价指标进行比较，分析方案的优劣。

此定价模型依然会有较多评价指标，因此，需要进行灵敏度分析，找出影响最灵敏的指标。预期可以得到修改的定价函数并且定价方案效果较好。经过一系列分析，可知此模型是适宜的。

三、问题假设

1、假设此众包平台各任务模式一样
2、在问题二中，由于任务颗粒度指标难易量化，仅考虑任务难度指标
3、用各任务点一定半径内会员人数衡量会员与任务的距离关系，进一步衡量任务难度
4、假设地面为一理想平面，不考虑地理因素
5、假设任务颗粒度为 1 时，可以考虑不降价
6、假设后续问题中会员的分布不发生变化

四、符号说明

a_1：会员在各个样本区域的个数，并且除以区域面积作为会员密度指标
a_2：任务在各个样本区域的个数，并且除以区域面积作为任务密度指标
a_3：各地区的任务平均价格指标
a_4：各个样本区域内任务完成率作为任务完成率指标
x_1：任务点一定范围内会员数
$f_i(x)$：新的定价规则函数
μ：原始定价规则下的平均价格
$G(x)$：关于任务困难度的函数
R：以任务点为圆心的圆的半径，每一单位表示经纬度一度跨越的距离
d_{1j}：第j种定价方案下定价指标
d_{2j}：第j种定价方案下任务完成率
f_{ji}：第j种定价方案下每单的价格
n_j：第j种定价方案下任务的总数。
Z_j：基于定价指标和任务完成率的综合评价指标
x_2：任务颗粒度
x_3：预定限额方差指标值
x_4：不同地区的经济指标
Y：所有地区的平均经济水平

五、模型的建立与求解

5.1 问题一模型的建立与求解

为了分析定价规律以及任务未完成原因，需要对已有数据进行分析，考虑到定价因素主要与任务和会员的分布有关，任务未完成原因与任务和会员分布及平均价格有关，所以先作出分布图，观察图形特征；其次通过聚类分析划分地区，以各地区的任务密度、会员密度为自变量，任务平均价格为因变量，采用偏最小二乘法求解定价关系；以各地区的任务密度、会员密度和平均价格为自变量，任务完成率为因变量，拟合关系，分析造成任务未完成的原因。

5.1.1 问题一的模型的建立

（1）数据可视化

首先，根据附录中所给任务点和会员的经纬度以及任务完成情况的相关数据，作出分布图，如图 5.1.1

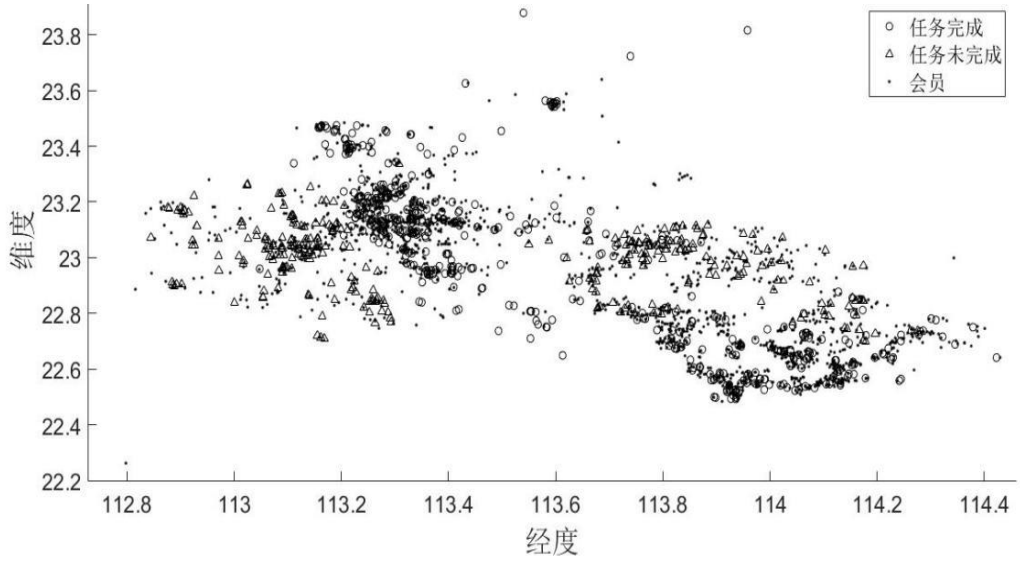

图 5.1.1 任务点和会员分布图

由图 5.1.1 可知，点表示拍照任务的分布位置，圆圈表示已完成的拍照任务点，方框表示未完成的任务点。一些任务集中地区会员也较为集中，但也存在任务较集中的地区会员较少较分散。

其次，根据任务与价格关系作图，如图 5.1.2

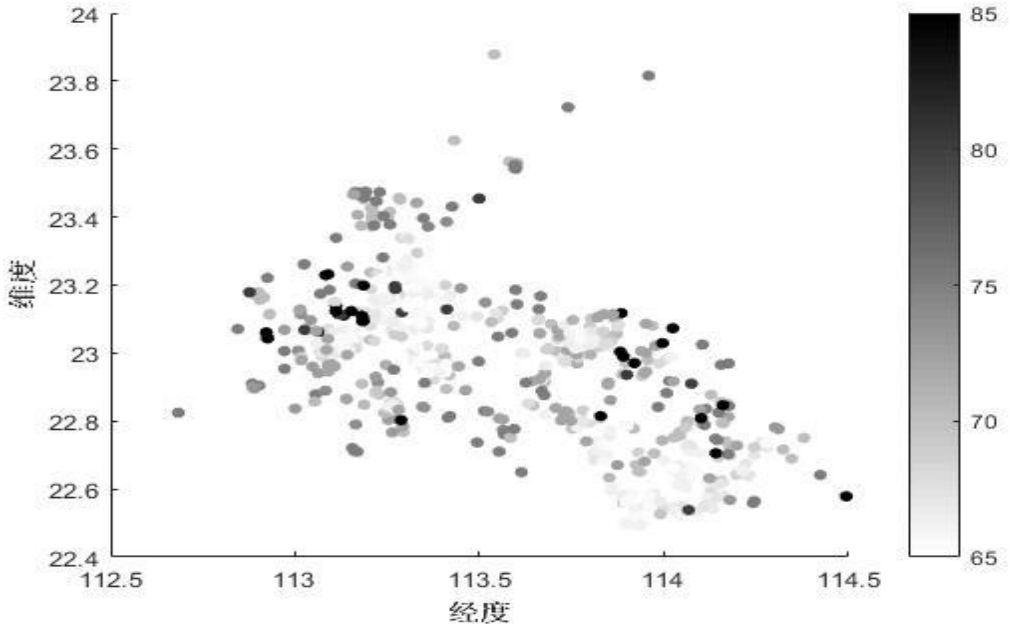

图 5.1.2 任务价格与分布点的关系图

图 5.1.2 中，根据任务点的经纬度确定分布位置，通过灰色度大小来表示价格高低，颜色由浅至深，分别表示价格从低到高。

（2）特征探索

将图 5.1.1 和图 5.1.2 进行对比，可以发现在会员较为集中的地区，任务价格普遍

偏低；而会员较为分散的地区，任务普遍价格偏高。

人们是否会参加众包类性质活动，往往取决于会员的兴趣、会员的位置、信誉值、参考其信誉给出的任务开始预订时间和预订限额和任务所能带来的收益情况等，而其中任务完成情况与定价有直接关系。

通过数据分析和实际情况的考虑，定价可能与该地区会员密度和任务密度存在一定的关系，而任务完成率与该地区会员密度、任务密度和任务价格等有相关性。为了简化模型，忽略人们的兴趣等影响。

（3）基于遗传模拟退火算法的聚类分析

聚类分析是研究多要素事物分类问题的数量方法。基本原理是根据样本自身的属性，用数学方法按照某种相似性或差异性指标，定量地确定样本之间的亲疏关系，并按这种亲疏关系程度对样本进行聚类。

各个任务点和会员位置这些个体所属类别和性质是未知的，但根据对相关数据的分析，这些样本点之间有一定的关系，为了更好地分析定价与任务和会员分布间的关系，于是，对任务点运用聚类的方法进行分类研究。

考虑到不同聚类方法的聚类效果有一定差异，在此采用基于遗传模拟退火算法的聚类法，它将模拟退火算法与遗传算法相结合，用于聚类分析[1]。

所以，利用基于遗传模拟退火算法的聚类法对各任务点的经纬度进行聚类分析。

（4）基于偏最小二乘法的回归模型的建立

偏最小二乘回归提供一种多对多线性回归建模的方法，特别当两组变量的个数很多，且都存在多重相关性，用偏最小二乘回归建立的模型具有传统的经典回归分析等方法所没有的优点[2]。

1、指标确立

采集会员在各个样本区域的个数，并且除以区域面积作为会员密度指标a_{1i}；采集任务在各个样本区域的个数，并且除以区域面积作为任务密度指标a_{2i}；采集任务的平均价格指标a_{3i}；采集各个样本区域内任务完成率作为任务完成率指标a_{4i}；。

2、基于偏最小二乘法的回归模型

偏最小二乘回归模型中集中了主成分分析和线性回归等分析方法的特点，因此在分析结果中更能提供一些更深入的信息。

在此，把各地区的任务密度、会员密度作为自变量，任务平均价格作为因变量；以任务密度、会员密度和平均价格作为自变量，任务完成率作为因变量，分别建立基于偏最小二乘法的回归模型。

对于p个因变量b_1，b_2，…，b_p与m个自变量a_1，a，…，a_p的建模问题，偏最小二乘回归的基本做法是：

首先，在自变量集中提出第一成分u_1（u_1是a_1，a_2，…，a_m的线性组合，且尽可能多地提取原自变量集中的变异信息）；同时在因变量集中也提取第一成分v_1，并要求第一成分u_1和v_1相关程度达到最大。

其次，建立因变量b_1，b_2，…，b_p与u_1的回归，继续第二对成分的提取，直到能达到满意的精度为止。

最终，对自变量集提取r个成分u_1，u_2，…，u_r,偏最小二乘回归将通过建立b_1，b_2，…，b_p与u_1，u_2，…，u_r的回归式，再表示为b_1，b_2，…，b_p与原变量的回归方程式，即偏最小二乘回归方程式。

5.1.2 模型的求解

首先将模拟退火和遗传算法相结合对数据进行聚类分析，其过程为

（1） 初始化控制参数：给定种群个体大小、最大进化次数、交叉概率、变异概率、退火初始温度、温度冷却系数以及终止温度

（2） 随机初始化 4 个聚类中心，并生成初始种群，对每个聚类中心计算各样本的隶属度以及每个个体的适应度

（3） 设循环初始变量为 0，对群体实施选择、交叉和变异等遗传操作，对新个体计算 4 个聚类中心、各样本隶属度和个体的适应度。若新适应度值大，则新个体替换旧个体，否则以一定概率接受新个体。

（4） 若循环变量小于最大进化次数，循环变量加 1，转至步骤（3），否则转（5）

（5） 若此时温度小于终止温度，则返回全局最优解；否则，执行降温操作

利用基于遗传模拟退火算法的聚类法对各任务点的经纬度进行聚类分析，划分出 4 个地区，具体效果图如下:

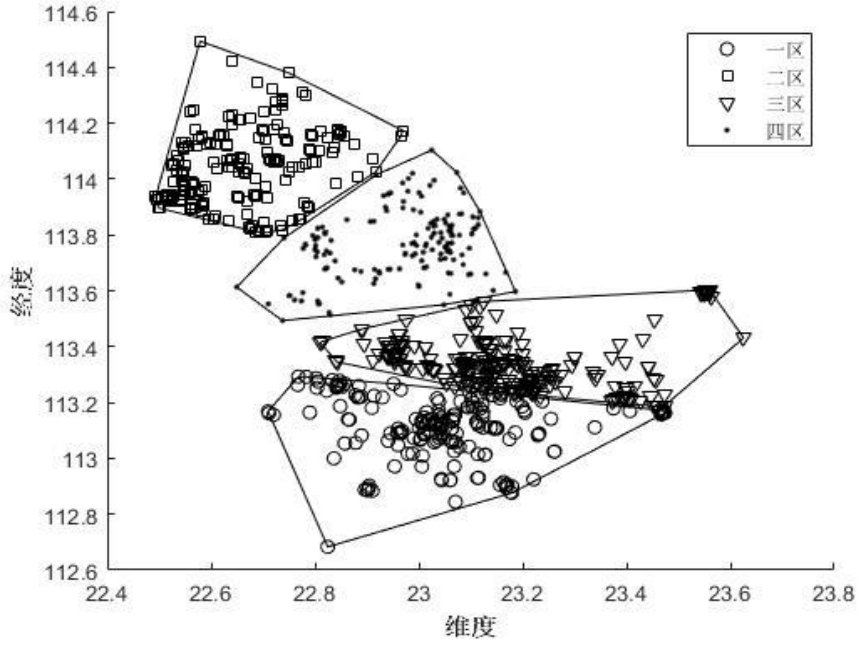

图 5.1.3 任务点的划分

如图 5.1.3，将各个任务点均划根据经纬度的差异进行聚类，分为四个不同的区域，分别表示为一区、二区、三区和四区。为了便于问题的研究，根据任务点聚类划区的情况，对每一地区采取一个样本圆区域作为本问题的分析区域，分为四个部分。

其次，建立基于偏最小二乘法的回归模型。

根据所求平均价格与会员密度、任务密度、信誉值、参考其信誉给出的任务开始预订时间和预订限额的关联，发现平均价格与会员密度和任务密度相关性较强；同理，任务完成率与会员密度、任务密度和平均价格相关性强。

于是，计算每个区域的会员密度、任务密度、任务完成率和平均价格指标如下：

表 5.1.1 不同区域各个指标值

区域	会员密度	任务密度	平均价格	任务完成率
一区	0.070	0.061	71.240	0.633
二区	0.130	0.068	70.080	0.987
三区	0.202	0.098	67.549	0.596
四区	0.256	0.077	68.297	0.349

分别将会员密度、任务密度与平均价格进行拟合，得到方程一，从而研究附件一中项目的任务定价规律；将会员密度、任务密度、平均价格与平任务密度进行拟合，得到方程二，由此分析任务未完成的原因。

为了消除不同变量量纲对总体拟合的影响，在进行偏最小二乘拟合之前，先根据不同的拟合方程对自变量进行标准化，得：

表 5.1.2 指标的标准化

会员密度	任务密度	平均价格
0.106	0.201	0.257
0.198	0.224	0.253
0.307	0.322	0.244
0.389	0.253	0.246

设自变量组和因变量组的n次标准化观测数据矩阵分别记为

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \quad B=\begin{bmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{np} \end{bmatrix}$$

设a_1、a_2和a_3分别为自变量会员密度、任务密度和平均价格，b_1为因变量平均价格，b_2为因变量任务完成率。

（1） 数据标准化

为了消除变量量纲对总体拟合的影响，先对自变量进行标准化。

将各指标a_{ij}转换成标准化指标值$\widetilde{a_{ij}}$，有

$$\widetilde{a_{ij}}=\frac{a_{ij}-\mu_j^{(1)}}{s_j^{(1)}}, i=1,2,\cdots,n, j=1,2,\cdots m$$

其中$\mu_j^{(1)}=\frac{1}{n}\sum_{i=1}^{n}a_{ij}$，$s_j^{(1)}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij}-\mu_j^{(1)})^2}$,即$\mu_j^{(1)}$和$s_j^{(1)}$为第$j$个自变量$x_j$的样本均值和样本标准差。

对应地，称

$$\widetilde{a_j}=\frac{a_j-\mu_j^{(1)}}{s_j^{(1)}}$$

为标准化指标变量。

类似地，将b_{ij}转换成标准化指标值$\widetilde{b_{ij}}$，有

$$\widetilde{b_{ij}}=\frac{b_{ij}-\mu_j^{(2)}}{s_j^{(2)}}, i=1,2,\cdots,n, j=1,2,\cdots m$$

其中$\mu_j^{(2)}=\frac{1}{n}\sum_{i=1}^{n}b_{ij}$，$s_j^{(2)}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(b_{ij}-\mu_j^{(2)})^2}$,即$\mu_j^{(2)}$和$s_j^{(2)}$为第$j$个自变量$y_j$的样本均值和样本标准差。

对应地，称

$$\widetilde{b_j}=\frac{b_j-\mu_j^{(2)}}{s_j^{(2)}}$$

为标准化指标变量。

（2） 求相关系数矩阵

表 5.1.3 各指标相关系数

相关系数	会员密度	任务密度	平均价格	任务完成率
会员密度	1	0.665	-0.897	-0.617
任务密度	0.665	1	-0.926	-0.310
平均价格	-0.897	-0.926	1	0.474
任务完成率	-0.617	-0.310	0.474	1

平均价格与会员密度和任务密度相关性较强；同理，任务完成率与会员密度、任务密度和平均价格相关性强[3]。

（3） 求因变量与自变量之间的回归方程

最终得到标准化指标变量之间的回归方程，将变量$\tilde{y_j}$和$\tilde{x_j}$分别还原成原始变量y_j和x_j,得到偏最小二乘回归方程为

方程一：$b = 75.689 - 6.832a_1 - 18.759a_2$

方程二：$b = -0.6044 - 1.4355a_1 + 1.2808a_2 + 5.1372a_3$

5.1.3 模型的检验

分别对回归方程一、方程二进行残差检验，得到四个样本的残差平方和分别为 0.0075、0.0032、0.6291、0.0438 和 6.51、2.73、402.63、37.21，方程一残差水平整体较小，拟合程度较好；而方程二的样本三偏差较大，其原因可能是区域三的样本数据点太过分散。

5.1.4 问题结果与分析

1、价格制定规律：

地区的平均价格与任务密度与会员密度的拟合方程为

$$b = 75.689 - 6.832a_1 - 18.759a_2$$

由此可见，平均价格与任务密度和会员密度成反比，任务密度和会员密度越大，则平均价格越低。任务点会员密度大，则会出现供小于求，从而使价格整体下降；若任务点密度大，任务点之间距离较小，减少了任务困难度，会大大引起会员的争抢，从平台角度考虑，不利于任务的完成和平台的效益，所以此时价格会降低。

2、任务完成率的影响因素：

地区的任务完成率与任务密度、会员密度和平均价格的拟合方程为

$$b = -0.6044 - 1.4355a_1 + 1.2808a_2 + 5.1372a_3$$

由此可知，任务完成率与任务密度成反比，与会员密度和平均价格成正比。该地区会员越多，同样供大于求，会提高任务的完成率；平均价格越高，提高了人们的兴趣，任务完成率自然提高；任务密度越大，供大于求，并且会引起人们对其的挑选与不重视，任务完成率便会下降。

5.1.5 问题小结

针对问题一，建立了基于偏最小二乘法的回归模型，研究了价格制定的规律，并分析了影响任务完成率的主要原因。首先，通过对问题的研究和数据的可视化，将任务点按经纬度差异性进行基于遗传模拟退火算法的聚类方法的分析，分为四个区域，确定并求出每个区域的任务密度、会员密度、任务完成率和平均价格等指标，利用基于偏最小二乘法的回归模型得到价格与任务密度和会员密度的关系、任务完成率与任务密度会员密度和价格的关系，最后对模型运用残差分析，验证模型合理性。

求解发现平均价格与任务密度和会员密度成反比；任务完成率与任务密度成反比，

与会员密度和平均价格成正比。

5.2 问题二模型的建立与求解

针对问题二，需要引入新的定价方案，并与原方案比较。首先，根据任务的难度指标，建立定价函数模型，确定新的定价方案；其次，利用极限学习机 ELM 对新方案进行仿真预测，并分析新方案下的任务完成情况；最后，建立基于定价指标以及任务完成率指标的综合评价模型，对两种方案进行优劣评价。

5.2.1 问题二模型的建立

（1）定价函数模型

对于这种众包性质任务，如何定价是目前研究的热点。一个任务的价格过高或者过低都会对任务的完成带来一定影响。如果价格过高可以吸引更多的工人来完成该任务，但是并不会使任务完成的质量有所提高，反而会增加任务请求人的经济负担，此外，任务价格过高容易引来欺诈者，导致任务的结果质量不高；而如果任务价格过低，则工人对此兴趣较低，导致任务很难被完成。

任务的难度和任务的颗粒是任务定价中两个重要的因素[4]，任务请求人需要综合这两方面因素，制定任务价格。同时，任务请求人在定价之前可以参照众包平台上类似任务的标价，该问题中任务颗粒度不好度量，于是仅依据任务难度指标建立定价模型。

由于任务的价格不应过高也不应过低[5]，考虑到对称性，一定存在一个基础价格，使实际价格在基础价格一定范围内波动，参照原始价格数据的平均值来构建基础价格，同时使用关于任务困难度的函数作为调节项，两者之和即是应定价格，即

$$f_1(x) = \mu + G(x_1)$$

式中，新的定价规则函数$f_1(x)$为关于任务点一定范围内会员数x_1的函数，μ为原始定价规则下的平均价格，$G(x_1)$为关于任务困难度的函数。

所以，在已知原始价格的平均值情况下，对于新的定价标准的确定，就转换为对于任务困难度调节项的研究分析。

1、任务难度指标的定量表现

任务难度取决于任务的模式以及任务消耗时长。由于此平台的任务全为拍照任务，所以假设各任务模式一样，那么任务难度可仅由任务消耗时长来衡量，而任务消耗时长仅根据任务地与会员的距离即可度量。

为了保证模型的简洁性，在此稍作变化，用各任务点一定半径内会员人数衡量任务难度，当范围内会员人数过少，即大部分人距离此任务较远，从而任务难度较高；当范围内会员人数过多，即大部分人距离此任务较近，从而任务难度较低。

所以，定义每个任务难度指标为x_1，表示在任务点半径为R范围内的会员人数。

2、调节函数的确定

当任务难度低时，可以适当减少价格；当任务难度高时，可以适当增加价格。考虑到任务难度指标最低为0时，减少的价格一定为一个有限值，同理，任务难度指标最大接近于+∞，此时增加的价格也应是一个有限值。假定增加或减少价格的最大值相等，于是可以列出基于神经元S特性的Sigmoid函数的调节函数，即

$$G(x_1) = \frac{1}{2}C_0 - \frac{C_0}{1 + e^{-C_1(x_1 - C_2)}}$$

令

$$g(x_1) = \frac{-C_0}{1 + e^{-C_1(x_1 - C_2)}}$$

式中，C_0、C_1、C_2均为正数。

所以，当$x_1 \in (0,+\infty)$时，有

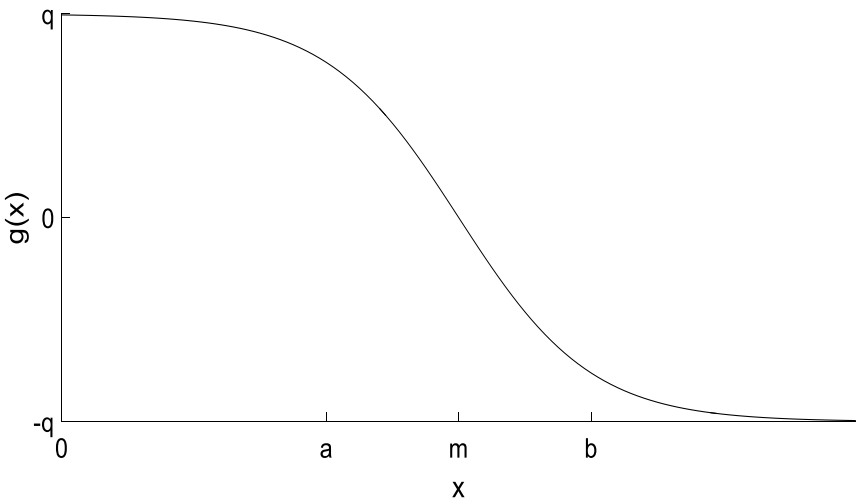

图5.2.1 基于神经元S特性的Sigmoid的函数

由图可知，当x_1为 0 时，$g(x_1)$接近于$\frac{C_0}{2}$，是一有界值q；当x_1趋于无穷时，此时$g(x_1)$无限趋于但不等于$-\frac{C_0}{2}$，为一有界值$-q$；当x_1为点$(m,0)$时图形关于该点对称，此时m取C_2，可见增加或减少的函数值是相等的；当x在[a,b]上时，x的变动对函数值影响较大，会员数在一定范围内变化时，价格变动大，超过了这个范围，价格变动则较小。所以，该图像与实际情况相符合。

之所以采用神经元 S 特性 Sigmoid 函数形式来描述调节函数，是因为 Sigmoid 函数具有连续、光滑、严格单调，且关于$(0,0.5)$中心对称等显著特点，因而在实践中 Sigmoid 函数往往作为优良的阈值函数。本文中采用该函数来描述关于任务困难度的调节函数，一方面可以保证调节值在一定范围内，另一方面也能描述调节函数与相关参数间的关系。

综上所述，可以得出基于Sigmoid函数的定价函数模型

$$f_1(x) = \mu + \frac{1}{2}C_0 - \frac{C_0}{1+e^{-C_1(x_1-C_2)}}$$

(2）基于极限学习机ELM的仿真预测模型

ELM 算法能深入挖掘既定数据中的隐藏信息，根据这些信息更好模拟出新的数据点。随机产生输入层与隐含层间的连接权值及隐含层神经元的阈值，且在训练过程中无需调整，只需设置隐含层神经元的个数，便可以获得唯一的最优解。

ELM 极限学习机模型在此问题中，不仅可以用来优化调节定价模型中的部分参数，还可以求解综合评价模型的任务完成率指标。通过 ELM 极限学习机进行机器学习，挖掘出隐藏在数据背后的灰色关系，良好地反映位置、价格对任务完成情况的影响，合理地对任务完成情况进行仿真分类预测[6]。

此建立基于极限学习机 ELM 的仿真预测模型，通过 ELM 学习算法，录入现有的 750 组数据，将其作为训练样本，将经、纬度和原先任务价格作为特征，是否完成任务作为分类结果，从而实现 ELM 的创建、训练和仿真测试，最后再次将样本经纬度和新的标准下各个任务的定价作为结果，预测出新规则下任务完成的情况。

典型的单隐含层前馈神经网络结构如图所示

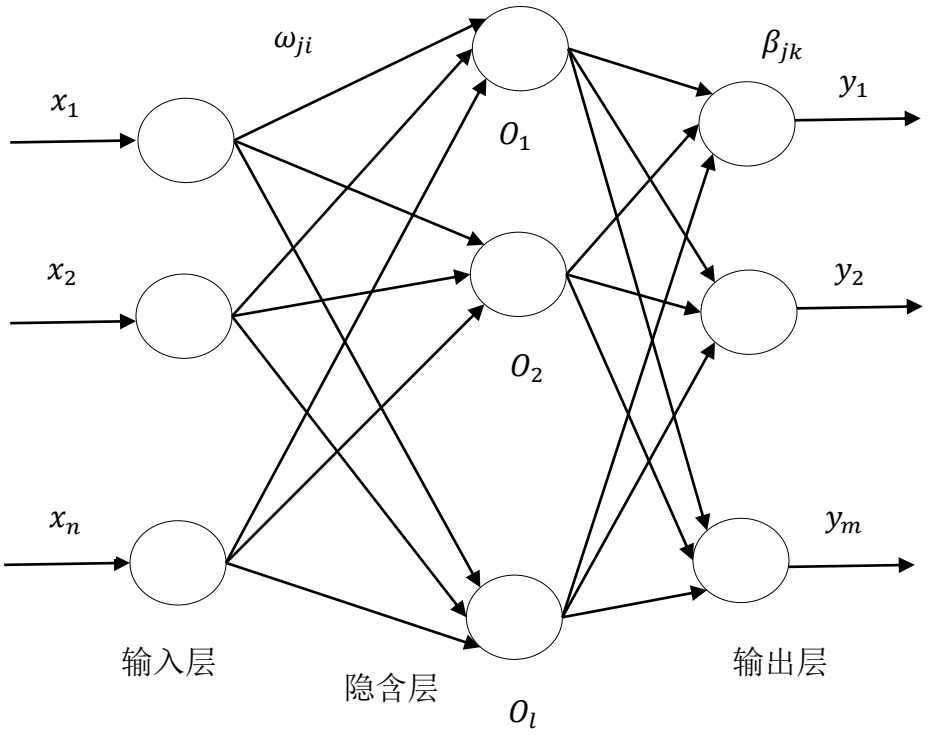

图 5.2.2 典型的单隐含层前馈神经网络结构

由输入层、隐含层和输出层组成，输入层与隐含层、隐含层与输出层神经元间全连接。其中，输入层有n个神经元，对应n个输入变量；隐含层有l个神经元；输入层有m个输出变量。为不失一般性，设输入层与隐含层间的连接权值w为

$$w = \begin{bmatrix} \omega_{11} & \cdots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{l1} & \cdots & \omega_{ln} \end{bmatrix}_{l\times n}$$

式中ω_{ji}表示输入层第i个神经元与隐含层第j个神经元间的连接权值。

设隐含层与输出层间的连接权值β和隐含层神经元的阈值b分别为

$$\beta = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1m} \\ \vdots & \ddots & \vdots \\ \beta_{l1} & \cdots & \beta_{lm} \end{bmatrix}_{l\times m} \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_l \end{bmatrix}_{l\times 1}$$

式中β_{jk}表示隐含层第j个神经元与输出层第k个神经元间的连接权值。

设具有Q个样本的训练集输入矩阵X和输出矩阵Y分别为

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1Q} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nQ} \end{bmatrix}_{n\times Q} \qquad Y = \begin{bmatrix} y & \cdots & y_{1Q} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nQ} \end{bmatrix}_{n\times Q}$$

设隐含层神经元的激活函数为$g(x)$，由单隐含层前馈神经网络结构图可知，网络的输出T为

$$T = [t_1, t_2, \cdots, t_Q]_{m\times Q}$$

而t_j为

$$t_j = \begin{bmatrix} t_{1j} \\ t_{2j} \\ \vdots \\ t_{mj} \end{bmatrix}_{m\times 1} = \begin{bmatrix} \sum_{i=1}^{l} \beta_{i1} g(w_i x_j + b_i) \\ \sum_{i=1}^{l} \beta_{i2} g(w_i x_j + b_i) \\ \vdots \\ \sum_{i=1}^{l} \beta_{im} g(w_i x_j + b_i) \end{bmatrix}_{m\times 1} \quad (j = 1,2,\cdots,Q)$$

式中$w_i = [\omega_{i1}, \omega_{i2}, \cdots, \omega_{in}]$，$x_j = [x_{1j}, x_{2j}, \cdots, x_{nj}]^T$

所以，网络的输出T的式子可以表示为

$$H\beta = T'$$

式中T'为矩阵T的转置，H称为神经网络的隐含层输出矩阵。

若隐含层神经元个数与训练集样本个数相等，则对于任意的w和b，SLFN 都可以零误差逼近训练样本，即

$$\sum_{j=1}^{Q} \left|\left| t_j - y_j \right|\right| = 0$$

式中$y_j = [y_{1j}, y_{2j}, \cdots, y_{mj}]^T (j = 1,2,\cdots,Q)$

然而，当训练集样本个数Q较大时，为了减少计算量，隐含层神经元个数通常取比Q小的个数，而又已知，SLFN 的训练误差可以逼近一个任意$\varepsilon > 0$，即

$$\sum_{j=1}^{Q} \left|\left| t_j - y_j \right|\right| < \varepsilon$$

因此当激活函数$g(x)$无限可微时，SLFN 的参数并不需要全部进行调整，w和b在训练前可以随机选择，且在训练过程中保持不变，而隐含层与输出层间的连接权值β可以通过求解以下方程组的最小二乘解获得

$$\min_{\beta} ||H\beta - T'||$$

其解为

$$\hat{\beta} = H^{+}T'$$

式中H^{+}为隐含层输出矩阵H的 Moore-Penrose 广义逆。

（3）基于定价指标以及任务完成率指标的综合评价模型

对两种定价方案进行综合评价，从成本方面提出定价指标，从收益方面提出任务完成率指标。

1、指标分析

定价指标是指在不同定价方案情况下的每一单任务平均价格，此指标属于成本型指标，即

$$d_{1j} = \frac{\sum_{i=1}^{n_j} f_{ji}}{n_j}$$

式中，f_{ji}表示第j种定价方案下每单的价格，n_j为第j种定价方案下任务的总数。

任务完成率是指在不同定价方案情况下，完成任务单数与发布单数的比值，属于效益型指标。值得注意的是，新定价方案的完成单数通过 ELM 神经网络进行仿真预测得出。所以，任务完成率表示为

$$d_{2j} = \frac{N_j}{n_j}$$

式中，N_j表示第j种定价方案下完成的任务单数。

2、指标预处理

对于成本型属性，令

$$d_{1j} = 1 - d_{1j} / \sum_{j=1}^{k} d_{1j}$$

对于效益型属性，令

$$d_{2j} = d_{2j} / \sum_{j=1}^{k} d_{2j}$$

显然，经过量纲化处理后，$d_j \in [0，1]$且d_{ij}的值越大越好。

定价指标和任务完成率指标根据其量化的定义，分别使用成本型和效益型方法进行去量纲处理。

3、指标权重确定

将定价指标看作投资风险，将任务完成率指标看作投资收益。实际投资中，不用的人由于财力、学识、时机等因素的不同，其投资风险的承受能力也不同，对于不同风险偏好的投资者其最佳投资疗法不同，为了反映实际情况我们将决策目标权重和投资风险偏好相联系，对应权重如下表，其中ω_1为投资收益的权重，ω_2为投资风险的权重

表 5.2.1 不同风险偏好的投资者的收益和风险权重

	高度冒险	比较冒险	中性冒险	比较保守	高度保守
ω_1	0.8	0.6	0.5	0.4	0.2
ω_2	0.2	0.4	0.5	0.6	0.8

对于众包类性质的 app 平台，它们通过开放集合互联网上的大众来完成一些任务，这些任务若按传统的工作方式进行，执行较为困难，成本较大，所以该平台的出现是为了节约成本高效完成任务，偏重于投资收益；但若任务定价即成本较低，会员会因此失去了参与的兴致，任务的完成率大大降低。从平台运营和会员角度出发，选择比较冒险的投资者收益和风险权重原则较好。

从而得到合理度综合指标：

$$Z_j = d_{1j}\omega_1 + d_{2j}\omega_2$$

比较两种方案下综合评价指标大小即可分析出新老方案的优劣程度。

5.2.2 模型的求解

1、参数R的确定

因为实际区域面积较小，所以在计算会员与任务地点之间的距离时，假设地面为一理想平面，直接采用欧氏距离计算以任务点为圆心，R为半径的圆内的会员数。根据距离关系，得到不同的任务点与其半径为R的圆内的会员数之间 835×1 的关系矩阵，其中行向量代表 835 个任务，列向量代表每个任务周围的会员人数。

从实际考虑，任务的位置不同，在不同任务点周围的会员人数也有较大偏差，导致其完成程度也会有较大差别。当需要比较几组数据离散程度大小的时候，如果几组数据的测量尺度相差太大，或者数据量纲的不同，直接使用标准差来进行比较不合适，此时就应当消除测量尺度和量纲的影响，所以，根据处理后的数据作出关系矩阵之间

的变异系数随R的变化图，如下

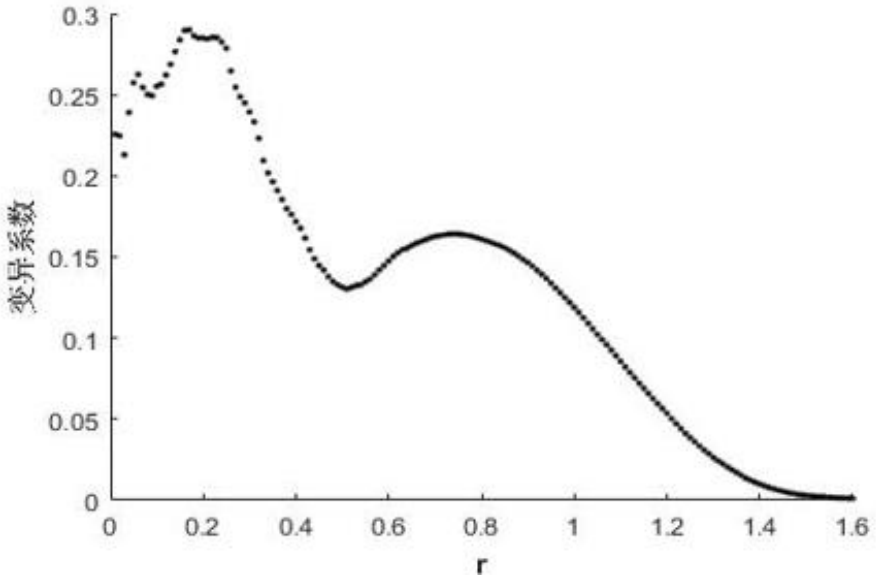

图 5.2.3 不同R与变异系数关系图

由上图可知，当R为 0.16,关系矩阵间的变异系数最大，此时任务点周围区域的划分最为明显，任务难度分布显著，便于对问题进一步的研究分析，所以，最终选取任务点半径为 0.16 时的不同任务点与其半径为R的圆内的会员数之间 835×1 的关系矩阵。

2、定价函数模型参数确定

首先得到原始定价价格的均值为 69 元（为了计算的简便，取整数），在新的定价标准中，C_0反映了价格的波动，当价格的增加使得原本未完成的任务完成，价格的减小使得原本完成的任务不再被完成。

为了确定C_0，在原始定价数据中随机选取完成任务点和未完成任务点各五组样本，改变原始价格，通过仿真模拟，得到任务类别发生改变（由 0 变为 1 或由 1 变为 0）时的最小价格变化，对十组数据取平均值即得到C_0=13.8。同时对曲线进行左右平移变化得到C_2=0.5。

C_1表示模型的变化趋势，对价格变化影响较大，将C_1视为自变量，对于每一个C_1，根据价格公式得到对应于不同任务的价格，其中，x_1表示不同任务点周围会员人数，为了消除x_1自身均值的影响，对其进行标准化处理。通过极限学习机 ELM 算法仿真模拟，得到不同任务点对应的完成情况。对于不同C_1所对应的任务完成情况和任务价格，分效益型指标和成本型指标进行数据预处理，权重分别为 0.6 与 0.4，得到不同C_1所对应的综合指标关系图，如下

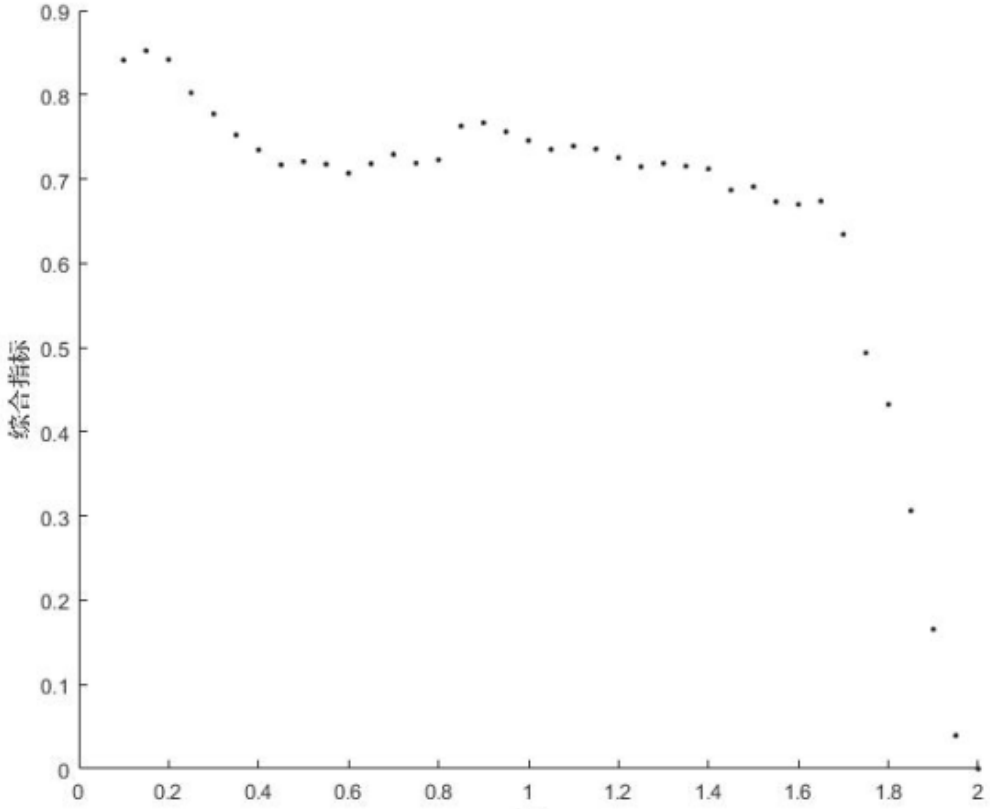

图 5.2.4 综合指标关系随C_1变化图

由关系图可知，当C_1=0.2 时，综合指标值取得最大值，此时新的价格模型最优，从而得到最终价格函数为

$$f_1(x) = 69 + 6.9 - \frac{13.8}{1 + e^{-0.2(x_1 - 0.5)}}$$

得到在两种定价标准下价格和任务完成率如下表所示：

表 5.2.2 新旧两种价格方案比较

价格方案	平均价格	任务完成率
原方案	69.11	0.63
新方案	69.02	0.79

根据综合评价公式

$$Z_j = d_{1j}\omega_1 + d_{2j}\omega_2$$

求得标准化之后的数据的综合指标为：

$$Z_1 = 0.46$$
$$Z_2 = 0.54$$

由此可见，新方案效果较原方案有较大提升，更好。

5.2.3 模型的检验

1、分析隐含层神经元个数对预测正确率的影响

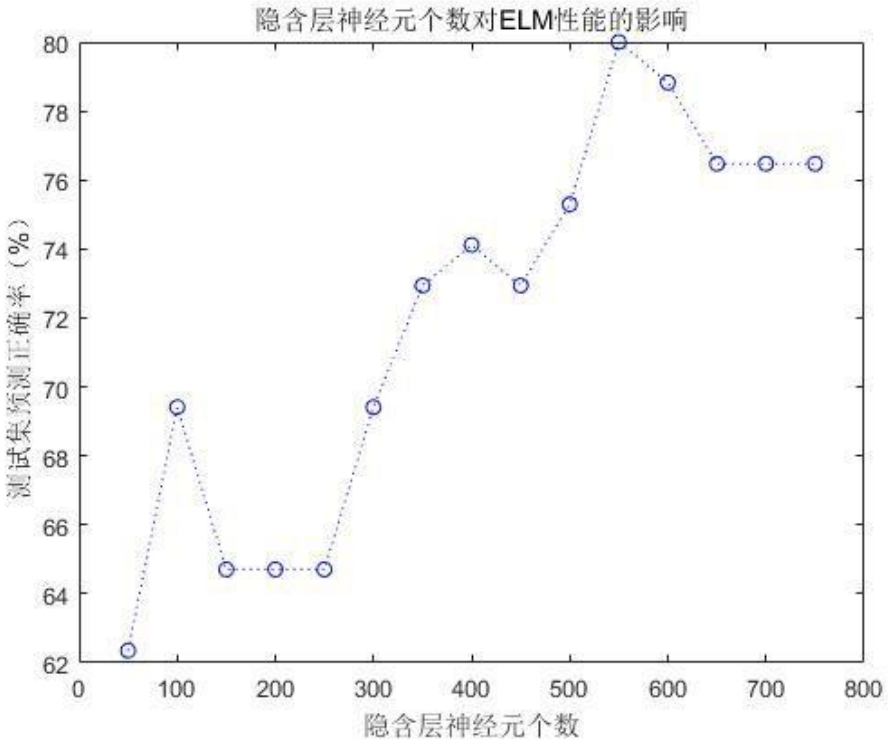

图 5.2.5 隐含层神经元个数对 ELM 性能的影响

由图 5.2.5 可知，并非隐含层神经元个数越多越好，从测试机的预测正确率可以看到，当隐含层神经元个数增加到 550 时，随着神经元个数的进一步增加，测试集的预测正确率并不会增加，相反还会减小。所以要综合考虑测试机的预测正确率和隐含层神经元的个数，进行折中选择。

2、指标的合理性分析

由两种定价标准下价格和任务完成率可知，在新的定价方案下，总体任务的完成率有显著提高，并且价格较原来也有一些降低，这就说明，新的定价方案实现了效益型指标的增长和成本型指标的降低，显然比原方案更好。

3、指标的灵敏性分析

以原始方案为例，保持权重不变，将完成率和价格指标分别增加 0.01，观察这两个指标的变化对综合指标的影响。实际计算表明，当完成率指标数值增加 0.01 时，综合指标增加了 0.57；当价格指标增加了 0.01 时，综合指标减小了 0.44，据此说明，完成率指标与综合指标成正比，价格指标与综合指标成反比，而完成率指标对总体指标的影响更大。

5.2.4 问题结果

通过求解，可知定价函数为

$$f_1(x) = 75.9 - \frac{13.8}{1 + e^{-0.2(x_1 - 0.5)}}$$

原定价规则和新的定价规则的综合评价指标分别为 0.46 和 0.54，由此可见新的定价规则更好。

完成率指标与综合指标成正比，价格指标与综合指标成反比，而完成率指标对综合指标的影响更大，该方案下总任务的完成率与原始情况下相比，提高较为明显，且价格也有一定的下降，利于平台的运营。

5.2.5 问题小结

针对问题二，建立了定价函数模型和基于定价指标以及任务完成率指标的综合评

价模型，确定新的定价方法并与原先方案进行优劣比较。根据众包类平台的运营特点，引入以任务点为中心的范围半径和任务困难度，在原始价格的基础上，列出基于神经元 S 特性的 Sigmoid 函数的调节函数，根据极限学习机 ELM 对原数据的仿真训练，预测出新方案下任务完成情况，并结合综合评价模型，确定定价函数的最优参数，并评价原始方案和新方案的好坏。最后，验证了模型预测结果和指标选取的合理性，对指标进行灵敏度分析发现完成率指标对综合指标影响较大。

通过求解，可知定价函数为

$$f_1(x) = 75.9 - \frac{13.8}{1 + e^{-0.2(x_1 - 0.5)}}$$

原定价规则和新的定价规则相比，新的定价规则更好，并实现了效益型指标的增长和成本型指标的降低。

5.3 问题三模型的建立与求解

针对问题三，要求考虑任务打包后，制定新的价格指标，并评价该方案的好坏。对打包现象提取新的指标变量，进而确定新的定价函数模型；其次，利用极限学习机 ELM 对新方案进行仿真预测，并分析新方案下的任务完成情况；最后，根据综合评价模型评价考虑打包方法后定价方案的优劣。

5.3.1 问题三模型的建立

（1）任务打包

考虑任务打包情况下的定价规则，应先将任务点按一定的规律进行分类，在此利用基于遗传模拟退火算法的聚类分析法。假设打包后每一类任务价格相等，将所有任务的平均位置作为该类的位置，从而分析每类任务的定价情况。

（2）定价函数的建立

对于任务的打包，在新的定价函数的基础上，引入的任务颗粒度和预定限额方差两个新指标，从而对原定价方案进行修改。

1、指标介绍以及定量化

任务颗粒度：指的是每个类里的任务数，任务间离散的程度，可以大致理解为与任务密度相关的一个值。任务颗粒度越大表示任务过于集中，因为任务距离较近会员可以接多个任务并且乐于去完成任务，可以达到做任务而不厌烦的一种状态。反之，任务颗粒度较小表示任务过于稀疏，会员做完一单后可能会不愿意再去较远的地方完成新的任务。

所以，任务颗粒度较大，人们乐于接单，此时可以适当降价，从而降低成本。

预定限额方差：会员的预定限额与信誉值有关并且大致能决定每个会员获得选择任务的几率，例如如果一个预定限额高的会员与预定限额低的会员同时竞争一单任务，那么很有可能是预定限额高的会员得到此单任务，在打包发布的时该因素影响较大。分析可知，这种机制可能会导致一种垄断效应，众包平台可能会一步步变为外包平台。

当任务一定区域内会员预定限额方差越大，代表此地区会员预定限额差距很大，价格应适当降低，否则易出现信誉高的会员垄断现象，不利于平台的发展。

2、定价函数的确定

在问题二定价函数基础上，再加上关于任务颗粒度和预定限额的调节项。该调节项任务颗粒度起主要作用，而预定限额方差则一定程度上控制了任务颗粒度对价格影响。

考虑当任务颗粒度即使趋近无穷，从平台角度和会员角度出发，其降价的额度也应该收敛到一个有效值，假设任务颗粒度为 1 时，可以考虑不降价。所以，可以将定

义为对数函数形式。

假定价格其他因素一定，横坐标表示任务颗粒度，纵坐标表示对任务价格的调节量，当任务颗粒度为 1 时，对价格没有影响；颗粒度由低开始变化时，对价格影响较大；当颗粒度趋于无穷时，任务价格趋于一定值。

当任务颗粒度一定，预定限额方差趋近于 0 时，此项指标值不对定价模型产生影响。所以，可以将其定义为指数函数形式。

假定价格其他因素一定，横坐标表示预定限额方差，纵坐标表示任务价格调节值，当预定限额方差越大，对价格影响越大，此时会造成会员垄断现象，与实际相符合。

综上考虑，在问题二的基础上，定义了新的函数项

$$h(x_2, x_3) = -C_3 a^{x_2} \ln(x_3)$$

式中，x_2、x_3分别表示任务颗粒度和预定限额方差指标值，C_3、a为常数。

所以，定价函数模型为

$$f_2(x) = \mu + \frac{1}{2}C_0 - \frac{C_0}{1 + e^{-C_1(x_1 - C_2)}} - C_3 a^{x_2} \ln(x_3)$$

（3）ELM 极限学习机仿真预测

使用 ELM 极限学习机以原数据为训练样本，根据经纬度以及新价格仿真预测任务完成情况，得到任务完成率指标，并且在此基础上调节模型的参数。具体实现过程参照问题二模型。

（4）综合评价模型

对定价方案进行综合评价，从成本方面提出定价指标，从收益方面提出任务完成率指标，将指标分别按成本属性和收益属性进行预处理，以比较冒险规则确定成本和利益指标的权重，最后得到综合评价指标

$$Z_j = d_{1j}\omega_1 + d_{2j}\omega_2$$

将打包后的综合指标与问题一问题二相比，分析打包能否有更好定价效果。

5.3.2 模型的求解

首先，将任务点按照经纬度通过聚类分析，将每一类中的任务点打包为一类。如图

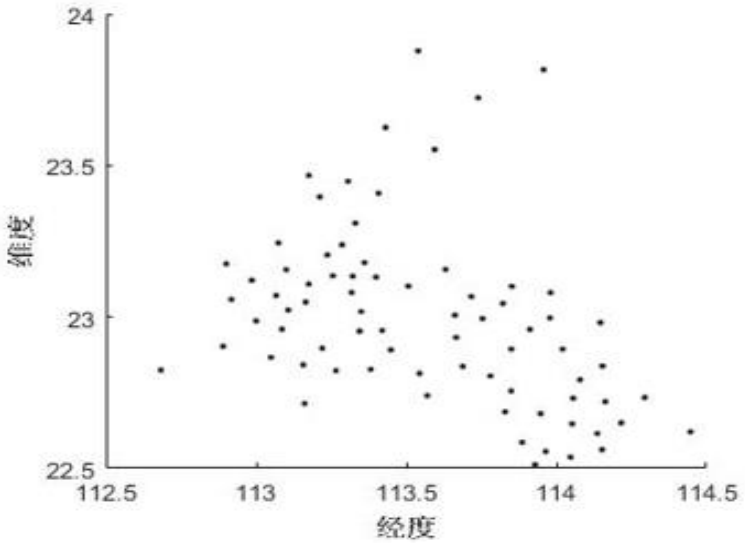

图 5.3.1 任务聚类打包图

图中每一点表示一类，将 835 个任务点打包为 75 个，以一类为一个样本点，研究定价函数模型。

对于模型的求解，主要在于对定价参数的确定。计算出每一类的各项指标值，如

任务颗粒度、预定限额方差指标值和会员数，在 ELM 学习机仿真模型和综合评价模型的基础上求出定价模型的参数。

为了确定C_3，在原始定价数据中随机选取完成任务点和未完成任务点，改变原始价格，通过仿真训练，模拟得到任务类别发生改变（由 0 变为 1 或由 1 变为 0）时的最小价格变化，对十组数据取平均值即得到C_3为 2。

a表示预定限额方差对价格的影响程度，先a视为自变量，对于每一个a，根据新的价格函数得到对应于不同任务类的价格，通过极限学习机 ELM 算法仿真模拟，得到不同任务类对应的完成情况。对于不同a所对应的任务完成情况和任务价格，分效益型指标和成本型指标进行数据预处理，权重分别为 0.6 与 0.4，得到不同a所对应的综合指标关系，从中选取评价指标最高情况下的参数值。发现当$a = 0.5$时，该定价方法最优。

所以 ，最终定价函数为

$$f_2(x) = 75.9 - \frac{13.8}{1 + e^{-0.2(x_1 - 0.5)}} - 0.5 \times 2^{x_2}\ln(x_3)$$

计算出该方案下所有任务平均价格为 60.879，已完成任务 683 个，

表 5.3.1 三种价格方案比较

价格方案	平均价格	任务完成率
方案一	69.11	0.63
方案二	69.02	0.79
方案三	60.88	0.82

根据公式

$$Z_j = d_{1j}\omega_1 + d_{2j}\omega_2$$

得综合指标为Z_1 =0.43、 Z_2 =0.47 、 Z_3 =0.50，可见打包后实施效果更好。

5.3.3 模型的检验

为了分析打包效果，随机选取一类，得到该类中任务点的分布图，如

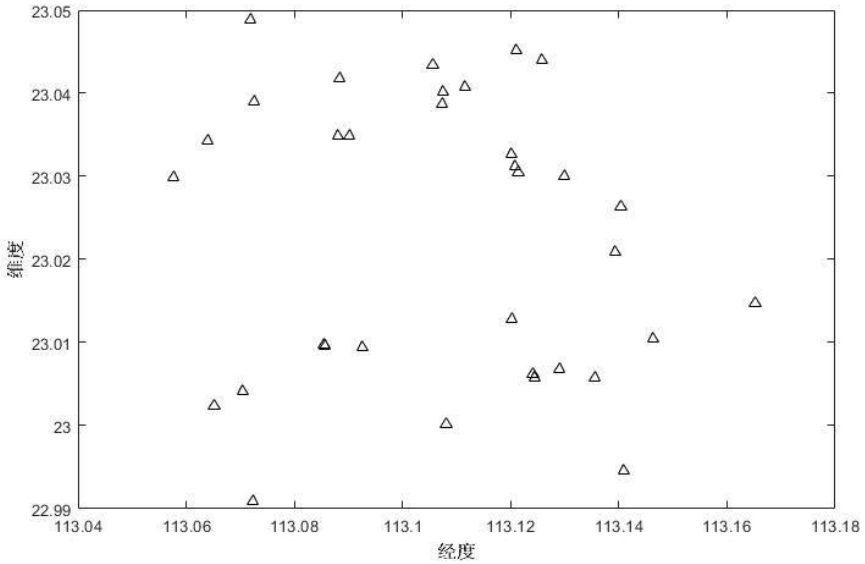

图 5.3.2 随机一类任务分布图

根据图可知，该范围内任务分布较为密集，均能在经度差为 2 度，纬度差为 1.5 度的长方形范围内，验证了聚类划分的合理性。

对模型中各指标进行灵敏度分析，发现当任务颗粒度、任务难度值和预定限额方差分别变化一个δ时，价格依次变化为 0.0345、 0.345 和 0.0352。所以，该定价模型中任务难度对价格影响最大。

5.3.4 问题结果

考虑任务打包发布后，在模型二定价函数的基础上加上关于任务颗粒度和预定限额方差指标的调节项，得新的定价函数为

$$f_2(x) = 75.9 - \frac{13.8}{1 + e^{-0.2(x_1 - 0.5)}} - 0.5 \times 2^{x_2} \ln(x_3)$$

其中，任务难度对价格影响最大。在这种方案下，任务完成率提高到0.82，与前两种方案相比，打包发布方案更优。

5.3.5 问题小结

针对问题三，同样建立了基于任务难度、任务颗粒度和预定限额方差的定价函数模型，并利用 ELM 学习机算法模拟预测模型和综合评价模型求出定价函数的参数，进而确定新的任务价格。

根据模拟预测的新价格下的任务完成情况，将综合指标与模型一、模型二中的相比，解决了考虑打包发布后任务价格的制定问题，并分析了完成情况的变化。最后验证了模型聚类结果的正确性，并对模型进行灵敏度分析，发现任务难度是影响价格的主要原因。

最后定价函数为

$$f_2(x) = 75.9 - \frac{13.8}{1 + e^{-0.2(x_1 - 0.5)}} - 0.5 \times 2^{x_2} \times \ln(x_3)$$

该方案任务完成率与前两种方案相比，有了明显提高，达到了 0.82；且综合评价指标值最大，定价规则最优。

5.4 问题四模型的建立求解

针对问题四，要求对附件三中的新项目给出你的任务定价方案，并评价该方案的实施效果。

在问题一和二的基础上分析研究，将地区进一步划分，分为广州和深圳，考虑不同市的会员对价格的期望的影响因素，确定各市的定价函数模型；通过利用极限学习机 ELM 对新方案进行仿真预测，并分析新方案下的任务完成情况；最后用综合评价模型评价方案实施效果。

5.4.1 问题四模型的建立

（1）地区经济指标的引入

根据附录三新任务的经纬度，画出它们的位置分布图 5.4.1。可知，新任务点的分布集中于两大块，依据对地图信息的查询[7]，可知集中区域分别为广州市和深圳市，不同地区会员对完成任务的所得期望不同，若任务价格的制定不考虑当地经济情况的影响，会导致会员对此失去兴趣，则不利于平台的运营，所以，引入地区经济指标，将其作为问题三中定价函数的调节项。

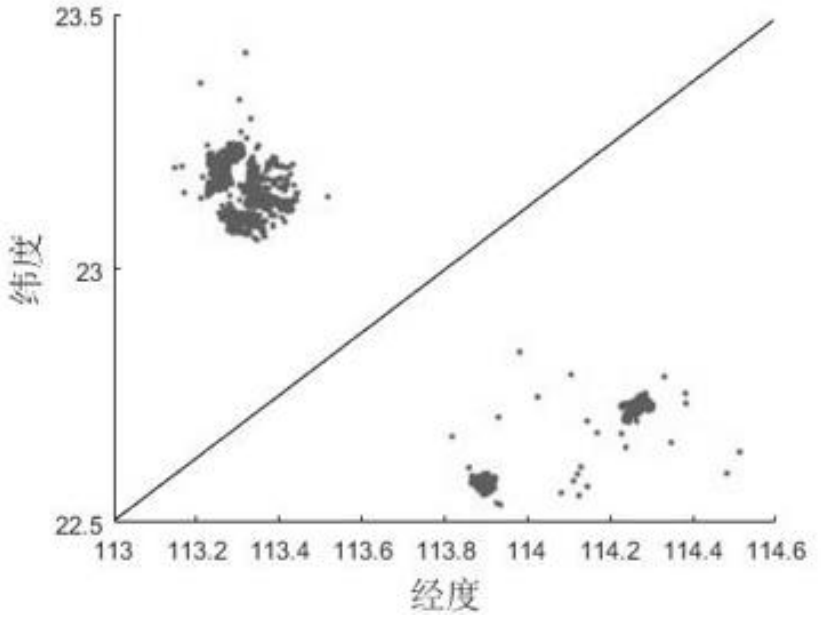

图 5.4.1 附录三任务的分布

（2）定价函数模型的建立

通过对统计的查询[8]，可知广州 2015 年国内生产总值为 18100.41 亿元，深圳 2015 年国内生产总值为 17502.86 亿元。由分析可知，经济指标与任务价格正相关，可得定价函数为

$$f_3(x) = \frac{C_4 x_4}{Y} f_2(x)$$

式中，x_4表示不同地区的经济指标，Y为所有地区的平均经济水平，C_4为常数，$f_2(x)$为模型三中的定价函数。

（3）ELM 极限学习机仿真预测

使用 ELM 极限学习机以原数据为训练样本，根据经纬度以及新价格仿真预测任务完成情况，得到任务完成率指标。具体实现过程参照问题二模型。

（4）综合评价模型

对定价方案进行综合评价，从成本方面提出定价指标，从收益方面提出任务完成率指标，将指标分别按成本属性和收益属性进行预处理，以比较冒险规则确定成本和利益指标的权重，最后得到综合评价指标

$$Z_j = d_{1j}\omega_1 + d_{2j}\omega_2$$

分析新定价方案下的实施效果。

5.4.2 模型的求解

首先，计算会员与任务地点之间的距离时，采用欧氏距离计算以任务点为圆心，R为半径的圆内的会员数。根据距离关系，得到不同的任务点与其半径为R的圆内的会员数之间的关系矩阵。从而得到处理后的数据与关系矩阵之间的变异系数随R的变化，计算发现，当关系矩阵间的变异系数最大时半径R为 0.025 时，此时任务点周围区域的划分最为明显。

为了确定该问题中定价模型的参数，利用学习机模拟预测模型和综合评价模型，从而求解较优方案下的参数C_4为 1.195。

所以，最终定价函数为

$$f_3(x) = \frac{1.195 x_4}{Y} f_2(x)$$

计算出该方案下任务的平均价格为 71.63 元，所有 2066 个任务中已完成任务为 1778 个，任务完成率达到 0.86，与问题一中方案 0.63 相比，提高较为显著。

根据综合评价模型中

$$Z_j = d_{1j}\omega_1 + d_{2j}\omega_2$$

求得问题一与该模型的综合评价指标分别为 0.46、0.54。

5.4.3 模型的检验

对模型所求得的定价函数进行灵敏度分析，当经济指标、任务难度、任务密度和预定限额方差分别变化一个较小的量δ时，任务价格依次变化为 0.034、0.351、0.035 和 0.035，表明任务难度对价格的变化最明显。

5.4.4 问题结果

该模型的最终定价函数为

$$f_3(x) = \frac{1.195x_4}{Y} f_2(x)$$

这种情况下，任务完成率达到了 0.86，平均价格为 71.63 元，与问题一的综合评价指标分别为 0.54 和 0.46，可见该模型更优。

5.4.5 问题小结

针对问题四，在问题三的基础上，对定价函数加以调整，引入地区经济变量，并利用 ELM 学习机算法模拟预测模型和综合评价模型求出任务划分范围半径和定价函数的参数，进而确定新的任务价格，解决了问题。最后对模型中定价函数进行灵敏度分析，发现任务难度仍是影响价格最重要的原因。

最后定价函数为

$$f_3(x) = \frac{1.195x_4}{Y} f_2(x)$$

任务完成率为 0.86，平均价格为 71.63 元，模型与问题一相比更合理。

六、模型的评价与改进

6.1 模型的评价

本文的亮点之一是指标合理的定义以及定量化。因为本文所选取的指标大部分都存在难以量化或者量化结果复杂的情况，我们对此类复杂指标进行侧面思考。例如本文的任务难度指标，结合这道题的背景，应该使用任务与会员的距离来定义这个指标，但是考虑到多个散点距离的求解复杂度，于是从另一个角度来衡量距离关系，根据任务点一定范围内会员的个数来度量。如果会员的距离离任务近，那么说明一定范围内会员数目较多；反之说明一定范围内会员数目较少。这样衡量方法，能确定任务难度指标的同时，可以使量化较为简单，编程易实现。同理，其余部分指标也做了相应处理，使模型结构变得简单明了。

本文的亮点之二是定价、ELM 仿真、综合评价三模型一体化的设计。最为重要的是 ELM 极限学习机模型的承上启下作用，它不仅可以用来优化调节定价模型中的部分参数，还可以求解综合评价模型中的任务完成率指标。通过 ELM 极限学习机进行机器学习，挖掘出隐藏在数据背后的灰色关系，良好地反映位置、价格对任务完成情况的影响。本文主要探讨的问题不仅是如何定价，更重要的是，无论定价函数是怎样的形式都能够良好地预测各种定价方案的优劣，有着良好的适应性。

本文的亮点之三是定价函数的确立。由于参考文献的缺乏，本文定价函数的确立主观因素较强，但是通过机理分析问题一中对任务完成情况的相关影响因素对合理定价的影响，可以找出性质相似且较好的函数作为关于各因素的价格模型。最终的仿真、

评价结果也反映出设计的新定价方案能够使得平均价格几乎一样，但任务完成率有明显提高，验证了定价函数的合理性。

6.2 模型的改进

本文提出三点模型改进方向

1、根据定价模型求得出来的价格属于固定价格，定价实现简单，但无法实现最小成本获得最大收益的目标，因为这种实现的前提需要综合考虑任务发布者与会员两方面因素。所以可以通过平台或者其他渠道按照时间，动态地搜集会员的信息，同时引入一种在线定价机制[9]，综合任务发布者与会员两方面确立最优价格。

2、本文的定价函数仅根据指标的机理性分析、函数优良性质和ELM极限学习机的参数优化建立得到，虽然依然能较好提升定价方案的效果，例如问题二中在价格指标差距不大时，可以明显提升任务完成率，但是不能保证这是最优的定价函数。

因此，类似于数理统计的参数估计中有效性的定义，给出方案有效性定义。只要找到最有效的定价函数，那么此定价函数即为最优的定价方案。

定义：

如果存在z_i、z_j并且满足条件

$$z_i > z_j$$

那么称$f_i(x)$定价方案比$f_j(x)$定价方案有效。其中z_i、z_j为第i、第j种定价方案的综合评价指标。

如果存在一个定价函数$f^*(x)$满足条件

$$z* \geq z$$

那么称$f^*(x)$为最有效定价函数。其中z表示任意方案的评价指标值。

3、综合评价模型也是贯穿全文的一个重要模型，由于仅从成本、收益方面定义了两个指标，所以一般仅通过主观方法赋权。但是实际上还可以找出更多的评价定价方案的指标，当指标数、数据量较大时，可以采用主客观结合的方法。例如建立基于层次分析法和熵权法的离差和最小的最优集成权重的综合评价模型[10]，对其进行综合评价，这样能最合理衡量定价方案的合理性。

参考文献

[1]武兆慧，张桂娟，刘希玉等 基于模拟退火遗传算法的聚类分析[J].计算机应用研究,2005,22(12):24-26.

[2]罗批，郭继昌，李锵等 基于偏最小二乘回归建模的探讨[J].天津大学学报(自然科学与工程技版),2002,35(6):783-786.

[3]莫莉，余新晓，赵阳等 北京市区域城市化程度与颗粒物污染的相关性分析[J].生态环境学报,2014,(5):806-811.

[4]冯剑红，李国良，冯建华等 众包技术研究综述[J].计算机学报2015,(9):1713-1726.

[5]王起功，郑富全，孙长青等 信息服务外包业产业化趋势及发展战略[C].//2012年山东省科协学术年会论文集.2012:1-11.

[6]甘露 极限学习机的研究与应用[D].西安电子科技大学,2014.

[7]中华人民共和国国家统计局，http://www.stats.gov.cn/，2017-09-17.

[8]高德地图经纬度查询，http://cloud.sinyway.com/Service/amap.html，2017-09-17.

[9]Yaron Singer,Manas Mittal.Pricing Mechanisms for Crowdsourcing Markets[C].//Proceedings of the 22nd international conference on world wide web, vol. 2: 22nd international conference on world wide web (WWW 13), May 13-17 2013, Rio de Janeiro, Brazil.2013:1157-1166.
[10]王中兴，李桥兴 依据主客观权重集成最终权重的一种方法[J].应用数学与计算数学学报,2006-06,20(1):87-91.

附录

附录只包含主要程序，详细程序见支撑材料

问题一

附录一：价格分布图

```
a=xlsread('C:\Users\Administrator\Desktop\附件一：已结束项目任务数据.xls','B2:B836');
b=xlsread('C:\Users\Administrator\Desktop\附件一：已结束项目任务数据.xls','C2:C836');
c=xlsread('C:\Users\Administrator\Desktop\附件一：已结束项目任务数据.xls','D2:D836');
scatter(b(:),a(:),25,c(:),'filled');
colorbar;
colormap(flipud(gray));
xlabel('经度','fontweight','bold');
ylabel('维度','fontweight','bold');
```

附录二：各区平均价格及会员关系

```
x=xlsread('C:\Users\Administrator\Desktop\chapter20\附件二：会员信息数据.xlsx','B2:B1878');
y=xlsread('C:\Users\Administrator\Desktop\chapter20\附件二：会员信息数据.xlsx','C2:C1878');
Z=xlsread('C:\Users\Administrator\Desktop\chapter20\附件二：会员信息数据新.xlsx','C2:C1878');
Y=xlsread('C:\Users\Administrator\Desktop\chapter20\附件一：已结束项目任务数据.xls','D2:D833');
%一区边界
xx =X(index1,1);
yy =X(index1,2);
pj1=mean(Y(index1))     %一区平均价格
k = convhull(xx,yy);
plot(xx(k),yy(k),'r-');
in(:,1)=inpolygon(x,y,(xx(k))',(yy(k))');%一区中所包含会员情况
z1=Z(find(in(:,1)==1),1);
vz1=var(z1)
hold on
%二区边界
xx =X(index2,1);
yy =X(index2,2);
pj2=mean(Y(index2))     %二区平均价格
k = convhull(xx,yy);
plot(xx(k),yy(k),'r-');
in(:,2)=inpolygon(x,y,(xx(k))',(yy(k))');%二区中所包含会员情况
z2=Z(find(in(:,2)==1),1);
vz2=var(z2)
%三区边界
xx =X(index3,1);
yy =X(index3,2);
pj3=mean(Y(index3))     %三区平均价格
k = convhull(xx,yy);
plot(xx(k),yy(k),'r-');
in(:,3)=inpolygon(x,y,(xx(k))',(yy(k))');%三区中所包含会员情况
z3=Z(find(in(:,3)==1),1);
vz3=var(z3)
%四区边界
xx =X(ial,1);
yy =X(ial,2);
pj4=mean(Y(ial))     %四区平均价格
k = convhull(xx,yy);
plot(xx(k),yy(k),'r-');
in(:,4)=inpolygon(x,y,(xx(k))',(yy(k))');%四区中所包含会员情况
z4=Z(find(in(:,4)==1),1);
```

```
vz4=var(z4)
fqsl=sum(in)%各区中所包含会员数量
```
附录三：聚类分析主函数（详细程序见支撑材料）
```
clc
clear all;close all
%load X %原程序数据调用方法，以下为实际应用调用方法
X=xlsread('C:\Users\Administrator\Desktop\chapter20\附件一：已结束项目任务数据.xls','B2:C836');
m=size(X,2);% 样本特征维数
% 中心点范围[lb;ub]
lb=min(X);
ub=max(X);
%% 模糊 C 均值聚类参数
% 设置幂指数为 3，最大迭代次数为 20，目标函数的终止容限为 1e-6
options=[3,20,1e-6];
% 类别数 cn
cn=4;
%% 模拟退火算法参数
q =0.8;        % 冷却系数
T0=100;       % 初始温度
Tend=1;   % 终止温度
%% 定义遗传算法参数
sizepop=10;                    %个体数目(Numbe of individuals)
MAXGEN=10;                  %最大遗传代数(Maximum number of generations)
NVAR=m*cn;                     %变量的维数
PRECI=10;                    %变量的二进制位数(Precision of variables)
GGAP=0.9;                    %代沟(Generation gap)
pc=0.7;
pm=0.01;
trace=zeros(NVAR+1,MAXGEN);
%建立区域描述器(Build field descriptor)
FieldD=[rep([PRECI],[1,NVAR]);rep([lb;ub],[1,cn]);rep([1;0;1;1],[1,NVAR])];
Chrom=crtbp(sizepop, NVAR*PRECI); % 创建初始种群
V=bs2rv(Chrom, FieldD);
ObjV=ObjFun(X,cn,V,options); %计算初始种群个体的目标函数值
T=T0;
while T>Tend
    gen=0;                                                    %代计数器
    while gen<MAXGEN                                          %迭代
        FitnV=ranking(ObjV);                                  %分配适应度值(Assign fitness
values)
        SelCh=select('sus', Chrom, FitnV, GGAP);              %选择
        SelCh=recombin('xovsp', SelCh,pc);                    %重组
        SelCh=mut(SelCh,pm);                                  %变异
        V=bs2rv(SelCh, FieldD);
        ObjVSel=ObjFun(X,cn,V,options);   %计算子代目标函数值
        [newChrom newObjV]=reins(Chrom, SelCh, 1, 1, ObjV, ObjVSel);        %重插入
        V=bs2rv(newChrom,FieldD);
        %是否替换旧个体
        for i=1:sizepop
            if ObjV(i)>newObjV(i)
                ObjV(i)=newObjV(i);
                Chrom(i,:)=newChrom(i,:);
            else
```

```
                    p=rand;
                    if p<=exp((newObjV(i)-ObjV(i))/T)
                        ObjV(i)=newObjV(i);
                        Chrom(i,:)=newChrom(i,:);
                    end
                end
            end
            gen=gen+1;   %代计数器增加
            [trace(end,gen),index]=min(ObjV);     %遗传算法性能跟踪
            trace(1:NVAR,gen)=V(index,:);%fprintf(1,'%d ',gen);
        end
        T=T*q;%fprintf(1,'\n 温度:%1.3f\n',T);
end
[newObjV,center,U]=ObjFun(X,cn,[trace(1:NVAR,end)]',options);   %计算最佳初始聚类中心的目标函数值
% 查看聚类结果
Jb=newObjV
U=U{1};
center=center{1}
figure
%% 分区
hold on
maxU = max(U);
index1 = find(U(1,:) == maxU);
index2 = find(U(2, :) == maxU);
index3 = find(U(3, :) == maxU);
% 在前三类样本数据中分别画上不同记号 不加记号的就是第四类了
plot(X(index1, 2),X(index1,1),'ko');
plot(X(index2, 2),X(index2,1),'ks');
plot(X(index3, 2),X(index3,1),'kv');
[al,ial]=setdiff(X(:,1),[X(index1,1);X(index2,1);X(index3,1)]);
plot(X(ial,2),X(ial,1),'k.');
xlabel('经度','fontweight','bold');
ylabel('维度','fontweight','bold');
%% 分区边界
xx =X(index1,1);
yy =X(index1,2);
k = convhull(xx,yy);
plot(yy(k),xx(k),'k-');%一区边界
area1=abs(trapz(xx(k),yy(k)))*111.11^2
%一区实际面积
xx =X(index2,1);
yy =X(index2,2);
k = convhull(xx,yy);
plot(yy(k),xx(k),'k-');%二区边界
area2=abs(trapz(xx(k),yy(k)))*111.11^2
%二区实际面积
xx =X(index3,1);
yy =X(index3,2);
k = convhull(xx,yy);
plot(yy(k),xx(k),'k-');%三区边界
area3=abs(trapz(xx(k),yy(k)))*111.11^2
%三区实际面积
xx =X(ial,1);
yy =X(ial,2);
```

```
k = convhull(xx,yy);
plot(yy(k),xx(k),'k-');%四区边界
area4=abs(trapz(xx(k),yy(k)))*111.11^2
%四区实际面积
legend('一区','二区','三区','四区'),hold off
```

附录四：拟合方程一（拟合方程二程序见支撑材料）

```
clc,clear
pz=xlsread('C:\Users\Administrator\Desktop\问题一指标数据.xlsx','H4:J7');
mu=mean(pz(:,1:3));
sig=std(pz(:,1:3));
%求均值和标准差
rr=corrcoef(pz(:,1:3));
%求相关系数矩阵
data=zscore(pz(:,1:3));
%数据标准化,变量记做 X*和 Y*
n=2;
m=1;
%n 是自变量的个数,m 是因变量的个数
x0=pz(:,1:n);y0=pz(:,n+1:end);
%原始的自变量和因变量数据
e0=data(:,1:n);
f0=data(:,n+1:end);
%标准化后的自变量和因变量数据
num=size(e0,1);
%求样本点的个数
chg=eye(n);
%w 到 w*变换矩阵的初始化
for i=1:n
%以下计算 w，w*和 t 的得分向量，
matrix=e0'*(f0*f0')*e0;
[vec,val]=eig(matrix);
%求特征值和特征向量
val=diag(val);
%提出对角线元素，即提出特征值
[val,ind]=sort(val,'descend');
w(:,i)=vec(:,ind(1));
%提出最大特征值对应的特征向量
w_star(:,i)=chg*w(:,i);
%计算 w*的取值
t(:,i)=e0*w(:,i);
%计算成分 ti 的得分
alpha=e0'*t(:,i)/(t(:,i)'*t(:,i));
%计算 alpha_i
chg=chg*(eye(n)-w(:,i)*alpha');
%计算 w 到 w*的变换矩阵
e=e0-t(:,i)*alpha';
%计算残差矩阵
e0=e;
%以下计算 ss(i)的值
beta=t\f0;
%求回归方程的系数，数据标准化，没有常数项
cancha=f0-t*beta;
%求残差矩阵
ss(i)=sum(sum(cancha.^2));
```

```
%求误差平方和
%以下计算 press(i)
for j=1:num
t1=t(:,1:i);f1=f0;
she_t=t1(j,:);she_f=f1(j,:);
%把舍去的第 j 个样本点保存起来
t1(j,:)=[];f1(j,:)=[];
%删除第 j 个观测值
beta1=[t1,ones(num-1,1)]\f1;
%求回归分析的系数,这里带有常数项
cancha=she_f-she_t*beta1(1:end-1,:)-beta1(end,:);
%求残差向量
press_i(j)=sum(cancha.^2);
%求误差平方和
end
press(i)=sum(press_i);
Q_h2(1)=1;
if i>1
    Q_h2(i)=1-press(i)/ss(i-1);
end
if Q_h2(i)<0.0975
    fprintf('提出的成分个数 r=%d',i); break
end
end
beta_z=t\f0;
%求 Y*关于 t 的回归系数
xishu=w_star*beta_z;
%求 Y*关于 X*的回归系数，每一列是一个回归方程
mu_x=mu(1:n);mu_y=mu(n+1:end);
%提出自变量和因变量的均值
sig_x=sig(1:n);sig_y=sig(n+1:end);
%提出自变量和因变量的标准差
ch0=mu_y-(mu_x./sig_x*xishu).*sig_y;
%计算原始数据回归方程的常数项
for i=1:m
xish(:,i)=xishu(:,i)./sig_x'*sig_y(i);
%计算原始数据回归方程的系数
end
sol=[ch0;xish]
%显示回归方程的系数，每一列是一个方程，每一列的第一个数是常数项
save mydata x0 y0 num xishu ch0 xish
```

问题二

附录五：求解包含原最佳半径（最佳半径时实际分布见支撑材料）

```
clc,clear;
rf=xlsread('C:\Users\Administrator\Desktop\附件一：已结束项目任务数据.xls','B2:C836');
%任务的位置坐标
hf=xlsread('C:\Users\Administrator\Desktop\附件二：会员信息数据.xlsx','B2:C1878');
%会员的位置坐标
p=1;
for k=0.01:0.01:1.6
    r=k;
    %任务圆的半径大小
    a=size(rf,1);
```

```
    b=size(hf,1);
    pd=zeros(a,b);
    for i=1:a
        for j=1:b
            rr=sqrt((rf(i,1)-hf(j,1))^2+(rf(i,2)-hf(j,2))^2);
            if rr<r
                pd(i,j)=1;
            end
        end
    end
    f=sum(pd,2);
    for t=1:835
        f(t)=(f(t)-min(f))/(max(f)-min(f));
    end
    y(p)=std(f);
    p=p+1;
end
x=0.01:0.01:1.6;
plot(x,y,'k.');
xlabel('r','fontweight','bold');
ylabel('变异系数','fontweight','bold');
附录六：确定参数
a=xlsread('C:\Users\Administrator\Desktop\附件二：会员信息数据.xlsx','H2:H836');
%分布数据标准化处理
k=1;%保证 k 为可引用矩阵下标，以下均作相同处理
for j=0.1:0.05:2
    for i=1:835
        a(i)=(a(i)-min(a))/(max(a)-min(a));
        f(i)=69-13.8/(1+exp(-j*(a(i)-0.5)))+6.9;
    end
    P_test1 = [data(1:835,5:6),f'];
    T_sim_3 = elmpredict(P_test1,IW,B,LW,TF,TYPE);
    ysws(k)=length(find(T_sim_3 == 2));
    %原始完成数
    ysjg(k)=mean(f);
    %原始价格均值
    k=k+1;
end
k=1;
for j=0.1:0.05:2
    bzhw(k)=(ysws(k)-min(ysws))/(max(ysws)-min(ysws));
    %将完成数视为效益型指标进行标准化
    bzhj(k)=(max(ysjg)-ysjg(k))/(max(ysjg)-min(ysjg));
    %将价格视为成本型指标进行标准化
    k=k+1;
end
kk=1;
for jj=0.1:0.05:2
    zh(kk)=0.4*bzhj(kk)+0.6*bzhw(kk);
    plot(jj,zh(kk),'k.');
    hold on
    kk=kk+1;
end
xlabel('C1','fontweight','bold');
ylabel('综合指标','fontweight','bold');
```

附录七：ELM 预测主函数（其余函数见支撑材料）

```
%% 极限学习机在分类问题中的应用研究
%% 清空环境变量
clear all
clc
warning off
%% 导入数据
data=xlsread('C:\Users\Administrator\Desktop\工作簿 3.xlsx');
data1=xlsread('C:\Users\Administrator\Desktop\工作簿 4.xlsx');
% 随机产生训练集/测试集
a = randperm(835);
Train = data(a(1:750),:);
Test = data(a(751:end),:);
% 训练数据
P_train = Train(:,1:3)';
T_train = Train(:,4)';
% 测试数据
P_test = Test(:,1:3)';
T_test = Test(:,4)';
P_test1 = data(:,5:7)';
fz=[22.54927013 114.1317447 73.5]';
tic
%% ELM 创建/训练
[IW,B,LW,TF,TYPE] = elmtrain(P_train,T_train,100,'sig',1);
%% ELM 仿真测试
T_sim_1 = elmpredict(P_train,IW,B,LW,TF,TYPE);
T_sim_2 = elmpredict(P_test,IW,B,LW,TF,TYPE);
T_sim_3 = elmpredict(P_test1,IW,B,LW,TF,TYPE);
T_sim_4 = elmpredict(fz,IW,B,LW,TF,TYPE);
toc
%% 结果对比
result_1 = [T_train' T_sim_1'];
result_2 = [T_test' T_sim_2'];
% 训练集正确率
k1 = length(find(T_train == T_sim_1));
n1 = length(T_train);
Accuracy_1 = k1 / n1 * 100;
disp(['训练集正确率 Accuracy = ' num2str(Accuracy_1) '%(' num2str(k1) '/' num2str(n1) ')'])
% 测试集正确率
k2 = length(find(T_test == T_sim_2));
n2 = length(T_test);
Accuracy_2 = k2 / n2 * 100;
disp(['测试集正确率 Accuracy = ' num2str(Accuracy_2) '%(' num2str(k2) '/' num2str(n2) ')'])
%% 显示
count_B = length(find(T_train == 2));
count_M = length(find(T_train == 1));
rate_B = count_B /750;
rate_M = count_M /750;
total_B = length(find(data(:,4) == 2));
total_M = length(find(data(:,4) == 1));
number_B = length(find(T_test == 2));
number_M = length(find(T_test == 1));
number_B_sim = length(find(T_sim_2 == 2 & T_test == 2));
number_M_sim = length(find(T_sim_2 == 1 & T_test == 1));
disp(['样本总数：' num2str(835)...
```

```
        '   完成：' num2str(total_B)...
        '   未完成：' num2str(total_M)]);
disp(['训练集病例总数：' num2str(750)...
        '   完成：' num2str(count_B)...
        '   未完成：' num2str(count_M)]);
disp(['测试集病例总数：' num2str(85)...
        '   良性：' num2str(number_B)...
        '   恶性：' num2str(number_M)]);
disp(['良性乳腺肿瘤确诊：' num2str(number_B_sim)...
        '   误诊：' num2str(number_B - number_B_sim)...
        '   确诊率 p1=' num2str(number_B_sim/number_B*100) '%']);
disp(['恶性乳腺肿瘤确诊：' num2str(number_M_sim)...
        '   误诊：' num2str(number_M - number_M_sim)...
        '   确诊率 p2=' num2str(number_M_sim/number_M*100) '%']);
R = [];
for i = 50:50:750
    %% ELM 创建/训练
    [IW,B,LW,TF,TYPE] = elmtrain(P_train,T_train,i,'sig',1);

    %% ELM 仿真测试
    T_sim_1 = elmpredict(P_train,IW,B,LW,TF,TYPE);
    T_sim_2 = elmpredict(P_test,IW,B,LW,TF,TYPE);
    %%  结果对比
    result_1 = [T_train' T_sim_1'];
    result_2 = [T_test' T_sim_2'];
    %  训练集正确率
    k1 = length(find(T_train == T_sim_1));
    n1 = length(T_train);
    Accuracy_1 = k1 / n1 * 100;
%       disp(['训练集正确率 Accuracy = ' num2str(Accuracy_1) '%(' num2str(k1) '/' num2str(n1) ')'])
    %  测试集正确率
    k2 = length(find(T_test == T_sim_2));
    n2 = length(T_test);
    Accuracy_2 = k2 / n2 * 100;
%       disp(['测试集正确率 Accuracy = ' num2str(Accuracy_2) '%(' num2str(k2) '/' num2str(n2) ')'])
    R = [R;Accuracy_1 Accuracy_2];
end
figure
plot(50:50:750,R(:,2),'b:o')
xlabel('隐含层神经元个数')
ylabel('测试集预测正确率（%）')
title('隐含层神经元个数对 ELM 性能的影响')
问题三
附录八：初步聚类（具体聚类程序见支撑材料）
a=xlsread('C:\Users\Administrator\Desktop\附件一：已结束项目任务数据.xls','B2:C833');
y=pdist(a,'seuclidean');
yc=squareform(y);
z=linkage(y,'ward');
dendrogram(z);
T=cluster(z,75);
for i=1:1:75
    tm=find(T==i);
    xx =max(a(tm,1))-min(a(tm,1));   %平均经度
```

```
    yy =max(a(tm,2))-min(a(tm,2));   %平均维度
    l=length(tm);%任务数量
    jzwa(i,1)=xx;
    jzwa(i,2)=yy;
    jzwa(i,3)=l;
end
```

附录九：会员分区情况（具体各区会员预定限额见支撑材料）

```
a=xlsread('C:\Users\Administrator\Desktop\附件一：已结束项目任务数据.xls','B2:C833');
y=pdist(a,'seuclidean');
yc=squareform(y);
z=linkage(y,'ward');
dendrogram(z);
T=cluster(z,75);
for i=1:1:75
    tm=find(T==i);
    xx =mean(a(tm,1));   %平均经度
    yy =mean(a(tm,2));   %平均维度
    l=length(tm);%任务数量
    r1=(max(a(tm,1))-min(a(tm,1)));%区域最大维度变化
    r2=(max(a(tm,2))-min(a(tm,2)));%区域最大经度变化
    r(i)=min(r1,r2)/2;%区域内接圆半径
    jzwa(i,1)=xx;
    jzwa(i,2)=yy;
    jzwa(i,3)=l;
    if i==53
        plot(a(tm,2),a(tm,1),'k^');
    end
    %具体判断分组点的集中情况
end
p=1;
for k=(mean(r)-0.1):0.001:(mean(r)+0.1)
    rr=k;
    %任务圆的半径大小
    aa=size(jzwa,1);%分类数
    bb=size(a,1);%任务点数
    pd=zeros(aa,bb);
    for i=1:aa
        for j=1:bb
            r_r=sqrt((jzwa(i,1)-a(j,1))^2+(jzwa(i,2)-a(j,2))^2);
            if r_r<rr
                pd(i,j)=1;
            end
        end
    end
    f=sum(pd,2);
    for t=1:75
        f(t)=(f(t)-min(f))/(max(f)-min(f));
    end
    vf(p)=std(f);
    p=p+1;
end
find(max(vf))
rx=(mean(r)-0.1):0.001:(mean(r)+0.1);
%plot(rx,vf,'k-');
%xlabel('r','fontweight','bold');
```

```
%ylabel('变异系数','fontweight','bold');
%% 不同区域内限额变异系数
hf=xlsread('C:\Users\Administrator\Desktop\附件二：会员信息数据.xlsx','B2:D1878');
%会员的位置坐标
r=0.16;
%任务圆的半径大小
a=75;
b=size(hf,1);
pd=zeros(a,b);
for i=1:a
    for j=1:b
        rr=sqrt((jzwa(i,1)-hf(j,1))^2+(jzwa(i,2)-hf(j,2))^2);
        if rr<r
            pd(i,j)=1;
        end
    end
end
for i=1:75
    ys=hf((find(pd(i,:)==1)),3);
    %每类原始限额
    tt=length(find(pd(i,:)==1));
    for t=1:tt
        bz(t)=(ys(t)-min(ys))/(max(ys)-min(ys));
        %数据分类标准化
    end
    jzwa(i,4)=std(bz);
end
附录十：ELM 预测
clc
%% 导入数据
data=xlsread('C:\Users\Administrator\Desktop\工作簿 3.xlsx');
data1=xlsread('C:\Users\Administrator\Desktop\工作簿 4.xlsx');
% 随机产生训练集/测试集
a = randperm(835);
Train = data(a(1:750),:);
Test = data(a(751:end),:);
% 训练数据
P_train = Train(:,1:3)';
T_train = Train(:,4)';
% 测试数据
X=xlsread('C:\Users\Administrator\Desktop\问题三\文件三数据表.xlsx','B2:F76');
x1=min(X(:,4));
x2=max(X(:,4));
for i=1:75
    X(i,4)=(X(i,4)-x1)/x2;
end
for j=1:75
    jg(j)=69+(6.9-13.8/(1+exp(-0.2*(X(j,5)-0.5))))+(-(1/2)*2^(X(j,4))*log(X(j,3)));
end
sj=ones(75,3);
sj(:,3)=jg';
sj(:,1:2)=X(:,1:2);
T_sim_5= elmpredict(sj',IW,B,LW,TF,TYPE);
length(find(T_sim_5==2));
(T_sim_5-1)*X(:,3)
```

```
plot(sj(:,2),sj(:,1),'k.');
xlabel('经度','fontweight','bold');
ylabel('维度','fontweight','bold');
```

问题四与问题三程序类似，不在此展示，具体见支撑材料