计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900024

基于树增益朴素贝叶斯网络的服务定价策略

韩丽霞[1]　张占营[2]
1 天津市西青区人民检察院　天津 300380
2 天津师范大学计算机与信息工程学院　天津 300387
(1984724553@qq.com)

摘　要　移动劳务众包是一种新型商业模式。服务定价问题是劳务众包平台的核心问题,它与任务完成度和企业利润密切相关。针对移动互联网中劳务众包平台的定价问题,对历史数据进行建模,探究定价和影响因素。采用多元线性回归对价格的主要影响因素进行拟合,研究了用户拍照任务执行情况、任务地理位置与拍照任务定价之间的函数关系。基于分治思想,使用树增益朴素贝叶斯网络(TAN)将地理信息划分为5个区域,将每个任务执行点用元组{任务完成度,任务标价,经度、纬度,信誉度}表示,对散点进行聚类分析,分析了任务未完成原因以及任务位置对任务完成情况的影响。提出区域会员信誉度计算方法,分别计算每个区域的信誉度,由信誉度和地理位置导出不同区域的价格,并评价该方案的实施效果。

关键词:树增益朴素贝叶斯网络;众包;定价策略
中图法分类号　TP399

TAN-based Service Pricing Strategy

HAN Li-xia[1] and ZHANG Zhan-ying[2]
1 The People's Procuratorate of Tianjin Xiqing District, Tianjin 300380, China
2 College of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China

Abstract　Aiming at the pricing problem of labor crowd sourcing platform in mobile Internet, this paper uses multiple linear regression to fit the main influencing factors of price. Based on the idea of divide and conquer, the geographic information is divided into five regions by using tree gain naive Bayesian network (TAN). Through clustering analysis, the scattered points are clustered, and the regional member reputation calculation method is proposed, and each region is calculated separately. According to the credibility and geographical location, the prices of different regions are derived. The solution proposed in this paper has a certain reference significance to the pricing problem, which is greatly affected by geographic information.

Keywords　TAN, Crowd sourcing, Pricing strategy

1　引言

众包是一种新型商业模式。在美国 *Wired* 2006年的6月刊上,记者 JeffHowe 首次推出了众包(crowd sourcing)的概念,即一个公司或机构把过去由员工执行的工作任务,以自由自愿的形式外包给非特定的大众网络的做法。众包描述了一种新的商业模式,即企业利用互联网来将工作分配出去,为企业提供各类信息搜集和商业检查,发现创意或解决问题。其最大优势就是通过网络控制,发挥志愿员工创意以及能力,使其愿意利用业余工作时间,满足于对其服务收取报酬。这对于企业来说,提供了一种组织劳动力的全新方式。

企业可以通过手机 APP 方式将所需要完成的任务发布,并且制定相应报酬,从而满足其市场检查、信息搜集等需求。全新的组织劳动方式,改变了传统市场调查方案,不仅能保证调查数据的真实性,也极大缩短了周期。这种方式大满足企业需求的同时,使得注册 APP 的会员也可利用自己的空闲时间赚钱。APP 就成为了此类劳务众包平台的核心。而任务定价又是其核心要素,定价问题与任务完成度和企业利润密切相关。任务定价太低,导致用户参与度和任务完成度过低;任务定价太高,则会增加任务发布者的成本。因此,合理定价对提高任务完成度、降低成本有着实际意义。定价问题,就是在运营商和会员的博弈中,寻找成本效益的最佳平衡点。定价策略对于当前劳务众包平台乃至互联网金融的发展具有重要意义。

随着众包模式的快速发展,移动众包的定价策略问题也备受关注[1],种类涵盖垄断平台定价[2]、竞争性平台定价[3]、动态定价[4]等。此外,也有一些学者研究了在线服务平台供需之间的有效匹配问题[5]。Wang 等[3]构建了考虑众包服务提供商之间竞争的动态定价模型,分析了竞争对最优服务价格的影响;Kung 和 Zhong[6]分析了基于会员的定价、基于交易的定价和交叉补贴3种定价策略。

上述研究虽然探讨了定价问题,但大多研究平台有统一定价策略,而根据移动 APP 服务外包的特点,任务执行点位置会影响用户活跃度。本文采用分治思想,自动分区并分区定价。

本文通过对历史数据进行建模,探究定价和影响因素,从而制定合理的任务定价方案。每个任务执行点由元组{任务完成度,任务标价,经度、纬度,信誉度}表示。

通信作者:张占营(zhanying999666@163.com)

首先，用贝叶斯网络将所处地域划分 5 个区域，建立初步模型，利用欧氏距离对模型进行检测，得出任务定价规律并且反思任务失败原因，筛选真实有效数据，参考信誉度问题得出全量表格，算出成本之后，重新根据不同位置定价并与原方案进行比较。

2 多元线性回归

假设影响定价的因素有地理位置和信誉度，采用多元线性回归进行分析。式(1)中，y 表示任务标价，x_1 表示纬度，x_2 表示经度，x_3 表示信誉度，β_0，β_1，β_2，β_k 为权重系数。

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k+\varepsilon \tag{1}$$

因素间互不相关，用最小二乘法最小化误差的平方和，得出参数值，如表 1 所列。

表 1 参数表

Table 1 Parameter

参数	x_1	x_2	x_3
回归系数	2.47679	0.27854	−19.43750
标准误差	0.80251	0.52786	72.73507
t Stat	3.08628	0.52768	−0.26723
P-value T	0.00209	0.59785	0.78935
下限 95.0%	0.90160	−0.75755	−162.20330
上限 95.0%	4.05199	1.31464	123.32829

得出关于定价的多元回归方程为：

$$y=2.47679x_1+0.27854x_2-19.43750x_3$$

每当任务点的纬度增加 1，任务标价增加 2.47679 元；每当任务经度增加 1，任务标价增加 0.27854 元。在回归统计表(见表 2)中，标准误差值较小，在表中 F 值为 0.001989，故置信度达到 99.8011。方程的拟合度较好，符合回归参数的显著性检验(t 检验)和回归方程的显著性检验(F 检验)。

表 2 回归统计表

Table 2 Regression statistical

参数	值
相关系数 R	0.121825
R^2	0.014841
调整后的 R^2	0.012473
标准误差	4.484539
自由度	2
误差平方和	252.07305
均方差	126.03652
F	0.001989

3 模型建立

本文基本思想是分而治之，利用树增益朴素贝叶斯、欧氏距离、聚类分析进行划片，将众多散点分区域管理。

贝叶斯公式用来描述两个条件概率之间的关系。贝叶斯网络(Bayesian Network，BN)，是一种对不确定性进行建模的有向无环图模型。BN 应用到分类模型中，便成为贝叶斯网络分类器(BNC)，朴素贝叶斯为其一种。它假设特征变量之间关于目标变量两两独立，网络结构独立。本研究中，定价的影响因素主要为地理因素与信誉度(通过回归线性方程也可以进行初步判断)，价格为目标变量，地理因素与信誉度作为特征变量。为了精准定义模型，采用树增益朴素贝叶斯网络(TAN)。寻找两种主要特征变量之间是否存在某种关系。

TAN 作为朴素贝叶斯网络的改进模型，假设属性对应的各个节点之间不一定相互独立。在 Z 已知的情况下，Y 可以提供给 X 的信息，用互信息表示如下。

$$I_p(X;Y|Z)=\sum_{x,y,z}P(x,y,z)\log\frac{P(x,y|z)}{P(x|z)P(y|z)} \tag{2}$$

TAN 保持了计算的复杂度和鲁棒性，具有更好的准确率。

聚类分析采用欧氏距离。设 $X_i=(x_{i1},x_{i2},x_{i3})$ 分别表示第 i 个数据的预订任务限额、预订任务开始时间、用户信誉值。设第 $j(j=1,2,3)$ 个指标统计值的样本均值为 μ_j，样本方差为 θ_j，进行分类之前应先进行标准化，做如下变换。

$$y_{ij}=\frac{x_{ij}-\mu_{ij}}{\theta_j} \tag{3}$$

设 d 为欧氏标准距离，即：

$$d_{ik}=\sum_{j=1}^{4}(y_{ij}-y_{kj})^2 \tag{4}$$

结果越小，表示情况越接近，可划分为一起。

4 模型应用与求解

4.1 区域划分

本文使用了“拍照赚钱”APP 任务运行的数据(任务所在经纬度、任务标价、任务是否完成)、会员基础信息。任务实验区域为广东的广州和深圳，共 835 个任务点(包含完成度、任务标价和 GPS 坐标)，1877 位会员信誉度信息。

利用 Matlab 建立散点图，如图 1 所示，依据散点密集情况，将散点划分为 5 个区域。横坐标为经度，纵坐标为纬度，蓝色为成功任务点，红色为失败任务点。并任务点 gps 参数定位到地图上，如图 2 所示。区域 1−5 的成功率为 0.32744，0.07751，0.16758，0.33939，0.0878。

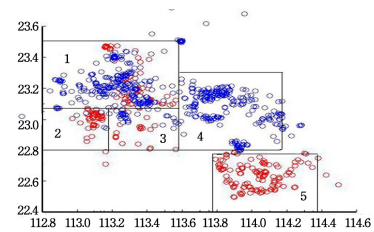

图 1 区域划分(电子版为彩色)

Fig. 1 Regionalization

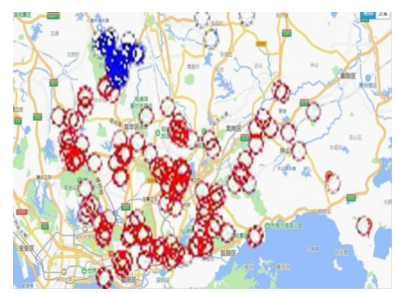

图 2 任务定位

Fig. 2 Task orientation

可以看出，散点分布具有地域聚集性特点，失败任务多集中在距离城市中心较远地区，说明任务成功率受地理因素影响较大。

从任务完成度可以看出，在会员信誉度高的区域，任务完成度相对较高。

4.2 完成定价

可以算出根据第一次任务经验得来的应有信誉度，即关键点数与贝叶斯成功率的积除以会员数。同时用关键点数与贝叶斯成功率的积除以任务量得到根据第一次任务经验得来的应有总钱数，将上面两个量做商得出的数为区域内一个信誉度对应的价格。需要在成本上增加的利润值为贝叶斯成功率与区域内一个信誉度对应价格的积再与根据第一次任务经验得来的应有信誉度作乘。将所有需要求出的量包括模型建立时的数据总结在一起。将信誉度顺序排序，如图 3 所示。

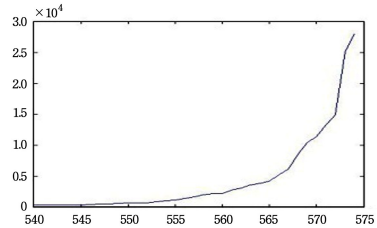

图 3 信誉度和会员数的关系

Fig. 3 Relationship between credibility and membership

假设会员数与整体区域会员信誉度之间存在某种特定关系。根据走势，设函数为：

$$y=ka^{bx} \tag{5}$$

其中，x 为会员数，y 为信誉度，k 为 x 处的斜率。

根据曲线走势，从 1 到 x 积分的值与从 x 到 574 积分的值相等，可得积分表达式如下：

$$\int_1^x ka^{bx}=\int_x^{574} ka^{bx} \tag{6}$$

$$\frac{ka^{bx}}{b\ln a}\int_1^x=\frac{ka^{bx}}{b\ln a}\int_x^{574}$$

$$a^{2bx}=a^{574b}+a^b$$

$$a^x=\sqrt{a^{574}+1}$$

$$x=\log_a\sqrt{a^{574}+1} \tag{7}$$

代入具体点(567，6 105.206)和(568，8 549.958)，将两个式子联立可得：

$$\begin{cases}6\,105.206=ka^{567b}\\8\,549.958=ka^{568b}\end{cases}$$

得出 $a\approx 0.19$。将 a 值代入式(7)，由于 0.19 的 574 次方趋近于 0，因此 x 趋近于 0，发现 $x=0$ 时的斜率即为成本。其他 4 个区域的关系图具有类似特点。

求得 5 个区域的定价分别为 72.0 元、69.7 元、65.9 元、70.4 元和 64.1 元。

以上主要参照信誉度，而对于地理条件，设价格为 y，距区域中心的距离为 x，且：

$$\Delta y=\frac{\Delta x}{100}$$

价格每增长 1，对应的距离增长 100 m，则区域 1 的定价方案为：

$$p_i=72.0+k\sqrt{(\text{经度}-113.3387)^2+(\text{纬度}-23.15934)^2}$$

其中，k 为经纬度距离和百米距离的转换系数。

各个区域使用的参数如表 3 所列，其中的参数依次为：平均价格 v_1，关键点数量 v_2，每个区域任务量比例 v_3，信誉度 v_4，会员数 v_5，贝叶斯成功率 v_6，任务完成度 v_7，任务量 v_8，根据第一次任务经验得来应有的信誉度 v_9，根据第一次任务经验得来应有总钱数 v_{10}，区域内一个信誉度对应的价格 v_{11}，需要在成本上增加的利润值 v_{12}。

表 3 数据表

Table 3 Data

参数	区域 1	区域 2	区域 3	区域 4	区域 5
v_1	68.846	69.739	68.642	70.410	67.253
v_2	14	0	1	5	4
v_3	0.34	0.113	0.131	0.208	0.208
v_4	301.201	20.579 6	15.243	20.904	39.622
v_5	574	109	142	340	619
v_6	0.327	0.076	0.168	0.0379	0.088
v_7	0.590	0.420 5	0.784	1	0.259
v_8	266	88	102	162	162
v_9	297.810	371.236	616.102	58.254	74.050
v_{10}	71.037	50.829	94.810	13.515	31.276
v_{11}	0.239	0.137	0.154	0.232	0.422
v_{12}	23.260	3.940	15.888	0.512	2.746

4.3 任务打包定价

采用多个任务打包发布的方式来提高任务的完成率并降低成本。首先需要确定打包的任务范围，然后给出打包任务定价规则。将数据所在地划分为若干区，利用 Matlab 得出每个区完成度与未完成度的散点分布图。设未完成任务为 x，以 x 为中心，查询半径为 r(据科学数据显示，成年人 10 min 大约可步行 1 500 m)的圆形区域的已完成任务数量 y，根据 x 和 y 的数量关系得出打包方案。用 C 语言编程，得出每一个区域对应方案，从而实现对每一个区域的分而治之。例如，对于区域 2 来说，这 106 个数据总共会产生的距离是 278.61 km，所以平均到每一个成功点到它所带的失败点的距离是278.61/106=2.628 km，并且区域 2 价格定价是 69.7 元，带两个任务之后，区域 2 每一点的价格=69.7+与之距离 2.68 km 失败任务点 1+与之距离 2.68 km 失败任务点 2。

结束语 本文旨在解决移动互联网自助劳务服务平台的任务定价问题。在多元线性回归分析之下，着重选取影响价格的重点因素，得出应着重考虑任务分布区域的地理条件以及各地区注册会员的信誉度。利于分治思想，首先将区域划分，初步建立模型，在还原成本基础上，充分利用地理信息以及会员信誉度重新制定价格，防范任务失败风险，同时提高利润值。本文提出的定价策略对受地理信息影响较大的定价问题具有一定的应用价值或参考意义。此外，本文尚未考虑未完成任务和行政区域划分，城市交通、人口密度、人均收入等社会因素的影响，以及任务分发方式、任务限时完成、任务难易程度等影响因素。另一个局限性是实验用的评估数据仅限于近几年，具有时效局限性。

参 考 文 献

[1] WANG S J，HOU Y. The application of regression analysis in the pricing of photo-earning task[J]. Journal of Guiyang University Natural Sciences，2019，14(1)：69-71.

[2] FENG Y Q，YAN L Y. A New Method of Crowdsourcing Platform Task Pricing[J]. Industrial Engineering and Management，2018，23(4)：145-149.

(下转第 219 页)

rithm:Algorithm AS 136 [J]. Applied Stats,1979,28(1):100-108.
[7] DING F,WANG J,GE J,et al. Anomaly detection in large-scale trajectories using hybrid grid-based hierarchical clustering[J]. International Journal of Robotics & Automation,2018,33(5):474-480.
[8] HUI F,PENG N,JING S,et al. Driving Behavior Clustering and Abnormal Detection Method Based on Agglomerative Hierarchy [J]. Computer Engineering,2018,44(12):196-201.
[9] MA M X,NGAN H,LIU W. Density-based Outlier Detection by Local Outlier Factor on Largescale Traffic Data[J]. Electronic Imaging,2016(2):385.
[10] WANG Y,PENG T,HAN J Y,et al. Density-Based Distributed Clustering Method[J]. Journal of Software,2017,28(11):2836-3850.
[11] LI N,QIANG Y,SUN Y,et al. Research on identification of aircraft abnormal trajectory in terminal area[J]. China Safety Science Journal(CSSJ),2018,28(11):21-27.
[12] LUXBURG U. A Tutorial on Spectral Clustering[J]. Statistics and Computing,2004,17:395-416.
[13] SHI J,MALIK J M. Normalized Cuts and Image Segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(8):888-904.
[14] NG A Y,JORDAN M I,WEISS Y. On Spectral Clustering:Analysis and an Algorithm[C]// Proceedings of the 14th International Conference on Neural Information Processing Systems:Natural and Synthetic. 2001:849-856.
[15] DU T T,WEN G Q,WU L,et al. Spectral clustering algorithm based on local covariance matrix[J]. Computer Engineering and Applications,2019,55(14):148-154,176.
[16] BHISSY K,FALEET F,ASHOUR W. Spectral Clustering Using Optimized Gaussian Kernel Function [J]. International Journal of Artificial Intelligence and Application for Smart Devices,2014,2:41-56.
[17] YU Q,LI Q,CHEN C,et al. Abnormal Trajectory Detection Method Based on BP Neural Network[J]. Computer Engineering,2019,45(7):229-236,241.
[18] TONG T,ZHU X,DU T. Connected graph decomposition for spectral clustering [J]. Multimedia Tools and Applications,2019,78(23).
[19] FANG M J,LIU M D. Similar measurement of time-space trajectory based on campus wireless network[J]. Computer Engineering and Design,2020,41(11):3001-3008.
[20] VLACHOS M,KOLLIOS G,GUNOPULOS D. Discovering similar multidimensional trajectories[C]// 18th International Conference on Data Engineering. IEEE,2002:673-684.
[21] PENG X,ZHANG L,YI Z. Scalable Sparse Subspace Clustering [C]// 2013 IEEE Conference on Computer Vision and Pattern Recognition. 2013:430-437.
[22] LI H,LIU J,LIU R W,et al. A Dimensionality Reduction-Based Multi-Step Clustering Method for Robust Vessel Trajectory Analysis[J]. Sensors,2017,17(8).
[23] ESTER M,KRIEGEL H-P,SANDER J,et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [C]// Proc. Int. Conf. Knowledg Discovery & Data Mining. 1996:226-231.
[24] WANG L J,DING S F,JIA H J. Spectral Clustering Algorithm Based on Message Passing[J]. Data Acquisition and Processing,2019,34(3):548-557.
[25] ROUSSEEUW P J. Silhouettes:A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics,1987,20:53-65.

GUO Yi-shan, born in 1997, postgraduate. Her main research interests include data mining in networks and intelligent optimization algorithm.

LIU Man-dan, born in 1973, Ph.D, professor, Ph.D supervisor. Her main research interests include control and optimization, application of intelligent methods, such as neural network and evolutionary computing, in control process.

(上接第 205 页)

[3] WANG W J,SUN Z M,XU Q. Dynamic Pricing for Crowdsourcing Logistics Services with Stochastic Demandand Competitive Providers [J]. Industrial Engineering and Management,2018,23(2):114-121.
[4] LIU W,YAN X,WEI W,et al. Pricing decisions for service platform with provider's threshold participating quantity, value-added service and matching ability[J]. Transportation Research Part E:Logistics and Transportation Review,2019,122:410-432.
[5] BAI J,SO K C,TANG C S,et al. Coordinating supply and demand on an on-demand service platform with impatient customers [J]. Manufacturing & Service Operations Management,2019,21(3):556-570.
[6] KUNG L C,ZHONG G Y. The optimal pricing strategy for two-sided platform delivery in the sharing economy[J]. Transportation Research Part E: Logistics and Transportation Review,2017,101:1-12.
[7] ZHU B X,MA Z Q,LI Z. Research on Incentive Mechanism of Performances of Cooperative Crowdsourcing Projects Based on Risk Prefence [J]. Industrial Engineering and Management,2019,24(3):60-68.
[8] DOU G,HE P,XU X. One-side value-added service investment and pricing strategies for a two-sided platform[J]. International Journal of Production Research,2016,54(13):3808-3821.

HAN Li-xia, born in 1985, postgraduate. Her main research interests include network security and so on.

ZHANG Zhan-ying, born in 1984, Ph.D, lecturer, is a member of China Computer Federation. His main research interests include iot and big data.