

Prediction of Success or Failure of Educational Crowds Based on Deep Neural Network

Liping Luo, Jie Huang, Xin Qin

PLA Strategic Support Force Information Engineering University, Zhengzhou Henan
Email: 47965213@qq.com, luopsa@163.com

Received: July 24th, 2019; accepted: August 8th, 2019; published: August 15th, 2019

Abstract

Educational crowdsourcing can optimize and test educational curricula to a certain extent, effectively integrate social and educational resources, and alleviate financial pressure for local governments. If crowdfunding fails, it will cause huge time costs. Therefore, it is of great significance to predict the results of crowdfunding projects. Aiming at the problem of predicting the success or failure of educational crowdfunding, this paper applies convolutional and neural network and BP neural network model to predict the success or failure of educational crowdfunding under the influence of multiple factors, and uses convolutional and neural network to train pure text information on the network; then introduces BP neural network, taking into account the influence of seven factors, such as text, total price and number of teachers of previously posted projects, and obtains the test accuracy of 89.3257%, and the impact of crowdsourcing forecast success or failure of the three main factors was analyzed.

Keywords

Educational Crowdfunding, Predicting the Success or Failure, Convolutional Neural Network, BP Neural Network

基于深度神经网络的教育众筹成败预测

骆丽萍, 黄洁, 秦鑫

中国人民解放军战略支援部队信息工程大学, 河南 郑州
Email: 47965213@qq.com, luopsa@163.com

收稿日期: 2019年7月24日; 录用日期: 2019年8月8日; 发布日期: 2019年8月15日

摘要

教育众筹在一定程度上可以优化、检验教育课程,有效整合社会资源及教育资源,为地方缓解资金压力,如果众筹失败,将造成巨大的时间成本,因此对众筹项目结果进行预测研究具有重要意义[1]。本文针对教育众筹的成败预测问题,将卷积神经网络和BP神经网络模型运用于多因素影响下的教育众筹成败预测中,利用卷积神经网络对纯文本信息进行网络训练;接着引入BP神经网络,综合考虑了文本、总价、以前公布项目的教师人数等七种因素的影响,获得了89.72%的测试正确率,并对影响众筹预测成败的三个主要因素进行了分析。

关键词

教育众筹, 成败预测, 卷积神经网络, BP神经网络

Copyright © 2019 by authors and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

众筹模式是在互联网不断创新发展下衍生出的一种新型网络融资模式,引入中国以来,迅速成为公众关注的热点[2]。教育众筹将众筹手段应用于教育领域,通过互联网方式帮助有需要的老师发布筹款项目并向网友募集资金,以此为提供更好的基础设施。一旦项目集资失败,会损失项目发起人和众筹参与者的时间,造成巨大的集资时间成本。因此,预测老师发布的请求书能否被社会人士认同并成功获得相应的教育资源成为了众筹平台和老师都迫切关注的问题。

目前国内关于众筹的研究大部分集中在众筹模式内涵、特征及法律风险等方面,对于众筹融资结果预测的研究比较少[3]。Lapedees (1987)团队首次将神经网络引入预测工作,国内对于神经网络的研究主要是BP (Back Propagation)神经网络研究。在BP神经网络的预测研究方面,有教学质量的预测评价模型,高速公路交通安全评价系统,对房地产市场进行预测研究等,发现预测误差较小[4]。BP算法能够帮助深度神经网络得到有效的训练,使得网络的结构可以向更深的方向发展。但当网络隐含层数量到达一定程度时,网络的性能往往不再提升,甚至会出现下降。卷积神经网络(Convolutional Neural Network, CNN)是近年发展起来并在计算机视觉和语音识别取得重大突破的一种深度神经网络,采用了局部连接和权值共享技术,不仅能够更好地提取特征信息,同时还减少了网络的参数,便于模型的训练[5]。

本文针对教育众筹的成败预测问题进行了深入研究,通过搜集的网上公开数据,在对自然语言处理的基础上,利用卷积神经网络和BP神经网络分析了文本和其他因素对教育众筹成败的影响,并给出了其中最重要的3个因素,以更好地指导教育众筹成功申请。

2. 基于深度神经网络的教育众筹成败预测

2.1. 深度神经网络理论

2.1.1. BP神经网络

深度学习概念源自于人工神经网络理论,Geoffrey Hinton指出,包含多个隐含层的人工网络具有更强的学习能力,更容易提取到数据的本质特征。依照层内相互独立,层间全连接的规则将人工神经元相

互连接,当隐含层超过 1 层时,我们就得到了一个简单的深度神经网络(Deep Neural Networks, DNN)。数据输入到神经网络后,在层内各个神经元中进行运算,输出结果通过全连接的方式输入到下一层中[6]。

如图 1 所示为一个三层 DNN,每层包含 3 个神经元。

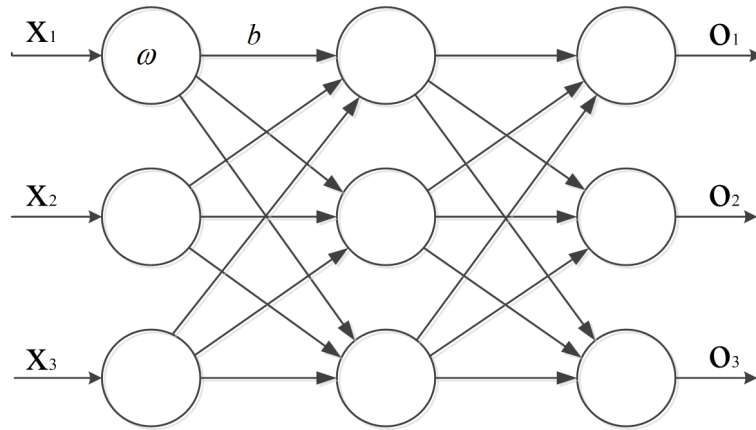


Figure 1. 3-Layer depth neural network
图 1. 3 层深度神经网络

设网络的输入为 $x_i, i=1,2,3$ 。假设没有激活函数,则第一层第 k 个神经元的输出为

$$o^{1k} = \sum_{i=1}^3 \omega_i^{1k} * x_i + b^{1k} \tag{1}$$

第二层第 l 个神经元输出为

$$o^{2l} = \sum_{k=1}^3 \omega_k^{2l} * o^{1k} + b^{2l} = \sum_{j=1}^3 \omega_j^{2l} * \left(\sum_{i=1}^3 \omega_i^{1k} * x_i + b^{1k} \right) + b^{2l} \tag{2}$$

显然,我们可以通过调整神经元的参数 ω 与偏置量 b 来使输出拟合我们想要的理想结果,调整优化参数的过程被称为神经网络的训练。

BP 算法的基本思想是利用求导的链式法则,在误差的反向传播过程中用梯度下降法来对神经元的参数进行调整。以图 1 中输出层的单个神经元为例,记:

$$z = \sum_{i=1}^m \omega_i * x_i + b \tag{3}$$

则神经元的输出为:

$$o = \sigma(z) \tag{4}$$

求神经网络损失函数 L 对 w_i 的偏导:

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial o} \frac{\partial o}{\partial z} \frac{\partial z}{\partial w_i} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial w_i} \tag{5}$$

显然 $\partial z / \partial w_i = x_i$, 在前向传播中已经得到 z 值,在激活函数确定的情况下可以算得 $\partial l / \partial z$ 。

进一步,对第 2 层即隐含层第 l 个神经元参数 w_i^{2l} 求偏导:

$$\frac{\partial L}{\partial w_i^{2l}} = \frac{\partial L}{\partial o^{2l}} \frac{\partial o^{2l}}{\partial z^{2l}} \frac{\partial z^{2l}}{\partial w_i} \tag{6}$$

因为神经元的输出为输出层的输入,所以

$$\frac{\partial l}{\partial o^{2l}} = \sum_{i=1}^3 \frac{\partial l}{\partial z^{3i}} \frac{\partial z^{3i}}{\partial o^{2l}} = \sum_{i=1}^3 \frac{\partial l}{\partial z^{3i}} x_i^{3i} \quad (7)$$

$\partial L/\partial z^{3i}$ 与 $\partial o^{2l}/\partial z^{2l}$ 已知,这样就求得了损失函数对前一层神经单元参数的偏导。继续向前依次递推,即可推得损失函数对所有神经元参数的偏导。进一步,就可以利用梯度下降法优化参数直至收敛。

2.1.2. 卷积神经网络

卷积神经网络是一种特殊的人工神经网络,目前已经成为语音分析和图像识别领域最常使用的工具之一。它是一个多层的神经网络,每层由多个二维平面组成,而每个平面由多个独立神经元组成,采用权值共享来减少网络参数规模[7]。

如图 2 所示,卷积神经网络由卷积层、池化层和全连接层组成,它通过对文本进行卷积和池化操作,能够从文本中提取更抽象的特征值以及单词的位置信息和单词间的相关语义信息,因而能够更好地用于文本分类[8] [9]。

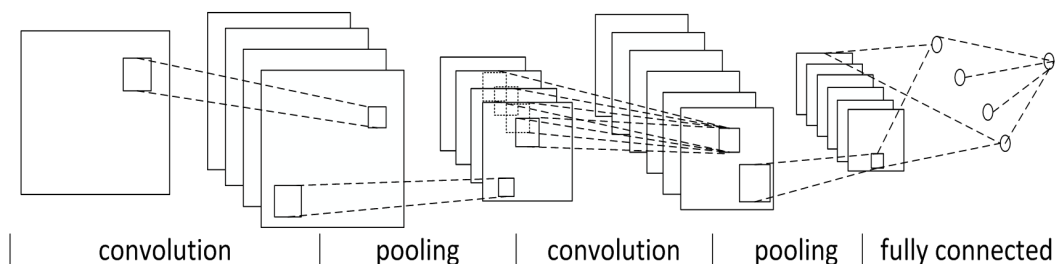


Figure 2. Structural sketch of convolutional neural network

图 2. 卷积神经网络结构示意图

卷积层本质上是一组滤波器,每个滤波器又称为卷积核。当图像输入网络时,单个卷积核通过固定步长下的不断滑动,与图像中的不同部分分别卷积,输出相应的特征二维矩阵。

卷积核的维度是三维的,其参数除了二维尺寸(Kernel Size),还有通道数(Channel)。当卷积层包含 n 个卷积核时,其输出 n 个二维特征图,此时下一层的卷积核的通道数维度应与上一层的输出相匹配。在参数总量固定的情况下,尺寸更小、卷积核数量更多的卷积层性能往往优于尺寸较大,卷积核数目较少的卷积层。卷积核如下式所示:

$$x_j^k = f \left(\left(\sum_{i \in M_j} x_i^{k-1} W_{ij}^k \right) + b_j^k \right) \quad (8)$$

式中, x_j^k 是第 k 层第 j 维的特征平面, M_j 是表示输入特征平面的集合, W_{ij}^k 表示由第 $k-1$ 层到第 k 层要产生的特征的数量,称为卷积核(Convolution Kernel)。卷积核可以看作一个四维矩阵,其中第一维是希望输出的特征平面数,第二维是当前层的特征平面数,第三、四维是局部感知域的大小。 b_j^k 表示偏置(Bias),是一个 k 维列向量, k 是输出的特征平面数。 $f(\cdot)$ 表示一个激活函数,本文使用的是 ReLU 函数。

池化是卷积神经网络另一种降低参数数目的手段,其本质是对特征的一种聚合操作。池化操作实际上是通过统计一定区域内的平均值(平均池化)或最大值(最大池化)实现降采样的作用。

全连接层即将人工神经元以层内独立、层间全连接的方式构造的网络结构。全连接层优势为结构易于调整,参数量少,其缺点在于丢弃了特征原本的空间结构。因此在 CNN 中,输入特征往往需要先经过多个卷积与池化的交替结构后,将特征高度抽象后,在通过全连接层展开并降维,以便最终通过 Softmax 函数输出分类结果。

2.2. 本文众筹预测模型

本文教育众筹成败的预测即是对文本分类的过程，预测模型由卷积神经网络和 BP 神经网络两部分组成。预测过程先对文本信息进行关键词提取、词频统计、词数值量化表征，通过卷积神经网络输出预测结果；再引入其余影响因素，数据形式的改变网络随之改变为 BP 神经网络，进行新一轮学习，如图 3 所示，对 BP 神经网络注入除文本信息以外的因素，重新学习，计算分类正确率。

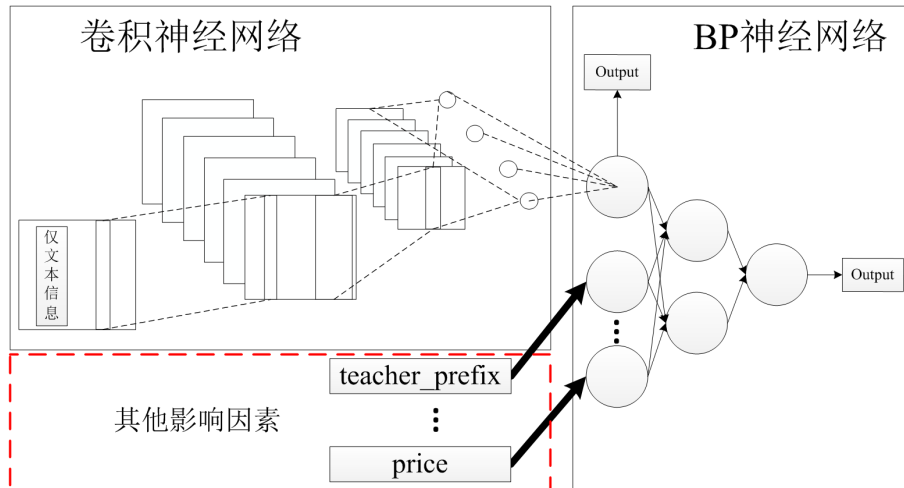


Figure 3. The schematic diagram of crowdsourcing forecasting principle in this paper
图 3. 本文众筹预测原理示意图

卷积神经网络卷积层的输入输出特征平面尺寸一般满足

$$x_{out} = \frac{x_{in} + 2pad - ks}{stride} + 1 \tag{9}$$

式中 pad 为填充宽度, ks 为卷积核尺寸, $stride$ 为步长。以 9×9 特征平面作为输入为例, 当 $pad = 0, ks = 3, stride = 1$ 时, 输出特征平面尺寸为 7×7 。

本文激活函数使用 ReLU 函数, 即

$$f(x) = \max(0, x) \tag{10}$$

2.3. 实验结果与分析

本文根据众筹网站 <http://www.donorschoose.org> 的数据进行仿真分析[10]。先对初始数据进行错误信息的剔除, 匹配等预处理后, 共获得 120597 个有效 id, 再对提取出的关键词进行数值量化处理。本文采用 Matlab2015b 作为软件平台, 编程实现 CNN 神经网络分类模型的构建、训练和分类, 将 50% 的 id 用于训练, 剩下的 50% 的 id 用于测试。

根据求得的词典和词语数值向量化的结果, 对每个文本进行数值矩阵化处理, 生成一个 $M \times 128$ 的特征矩阵, 其中 M 为本文中能够与词典中的词语匹配的词语数, 为了使每个文本的矩阵维数相同, 选取所有文本中最大的词语匹配数, 这里选取 $M = 200$ 。

利用特征矩阵, 对卷积神经网络进行训练, 本文设计了感知野为 $M \times 128$ 的长矩形卷积核, 保证每次卷积中单词信息的完整; 针对单词间关联距离不确定问题, 设计了 64 个 $M = 3/4/5$ 三种尺寸的卷积核, 同时对输入数据进行处理。不同卷积核的输出在融合层与全连接层进行信息融合与抽象, 从而取得较之单一卷积核尺寸更好地结果。最终交叉验证正确率可达 88.16%, 结果如表 1 所示。

Table 1. Comparison of network performance under different perceptions
表 1. 不同感知下网络性能比较

卷积核尺寸	3 × 128	4 × 128	5 × 128	本文网络
准确率	85.14%	85.31%	86.34%	88.16%

这里在文本信息的基础上，进一步考虑其他因素对分类器性能的影响，除文本信息外，加入跟教师相关、学校情况、项目提交日期、项目类别等级、提交项目的类别、以前公布项目的教师人数、总价等 7 个因素，采用 BP 神经网络，网络训练过程参数如图 4 所示。

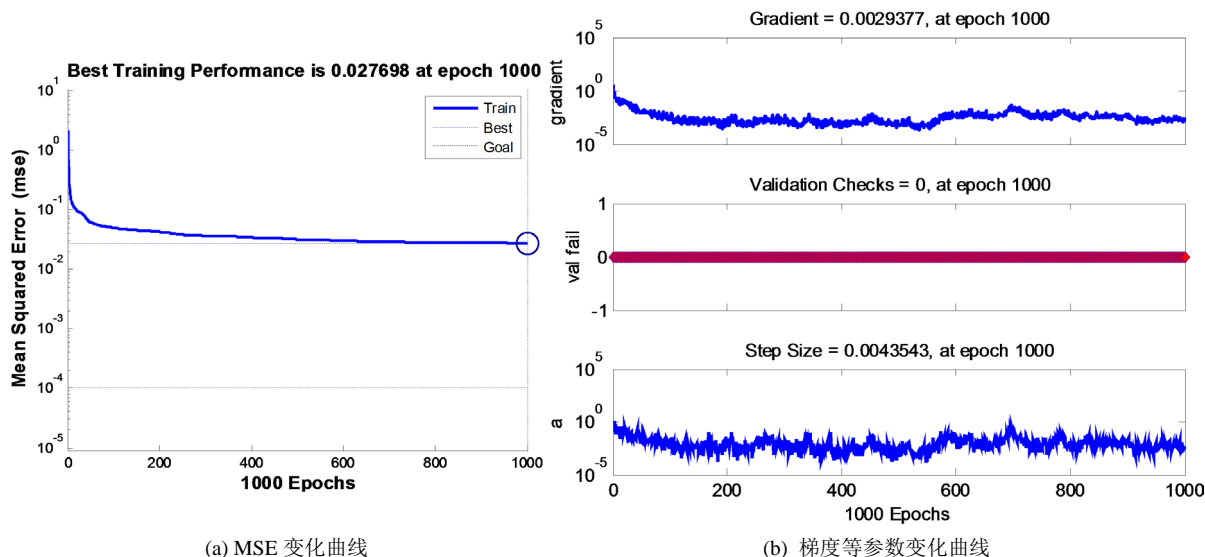


Figure 4. Variation of parameters in training process of BP neural network
图 4. BP 神经网络训练过程参数变化

利用上述神经网络对测试样本进行测试，共测试 10,000 个样本，这里仅给出 200 个样本预测结果和真实结果的对比图，如图 5 所示，可以看出，200 个样本中有 2 个样本原本为 approved 被判别为 not approved，而有 12 个样本原为 not approved 被分类为 approved，具有较高的正确率。10,000 个测试样本的正确率可达 89.72%。

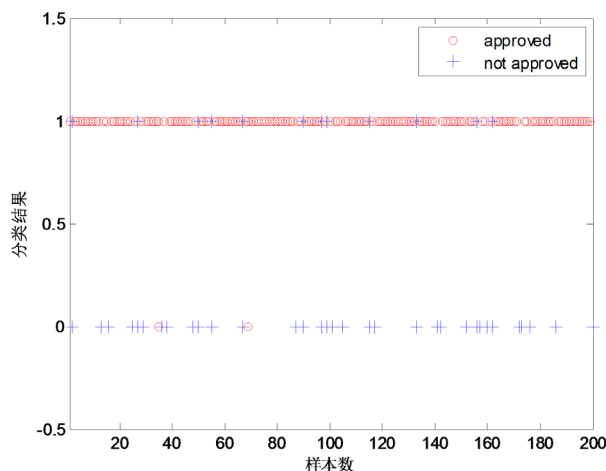


Figure 5. Test result diagram under 200 samples
图 5. 200 个样本下的测试结果图

3. 各因素对教育众筹成败预测的影响

为了给出每个因素对请求书的影响，这里采用因素递减对比的方法，即每次减去因素，利用剩余因素进行神经网络的训练和预测，并给相应的测试结果，如表 2 所示。

Table 2. Detection accuracy after factor decreases
表 2. 因素递减后的检测正确率

因素	去除相应因素后的正确率
跟教师相关	89.69%
学校情况	89.42%
项目提交日期	89.72%
项目类别等级	89.71%
提交项目的类别	89.56%
总价	89.17%
以前公布项目的教师人数	88.94%
文本	84.83%

由表可知对请求书影响最大的三个因素为文本、总价和以前公布项目的教师人数。为了进一步探索每个因素是如何对结果产生的影响，这里给出每个因素在不同取值下请求书被支持的成功率。图 6 给出了以前公布项目的教师人数对项目支持率的影响结果。图 6(a)为所给表格中不同以前公布项目的教师人数对应的总人数，由于以前公布项目的教师人数大于 100 后，其人数相对较少，不具有统计意义。因此，图 6(b)仅给出 0 到 100 范围内所对应的支持率，可以看出其总体趋势为以前公布项目的教师人数越大，支持率越高。

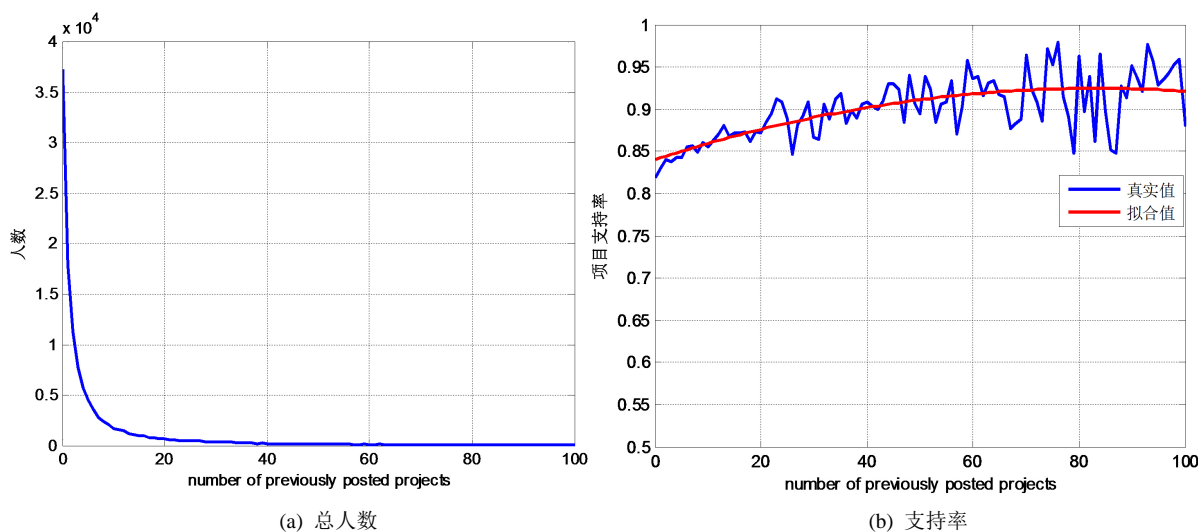


Figure 6. Impact on project support rate of teacher_number_of_previously_posted_projects
图 6. teacher_number_of_previously_posted_projects 对项目支持率的影响。

图 7 给出了总价对项目支持率的影响结果。图 7(a)为所给表格中不同总价对应的总人数，这里可以发现当总价大于 1000 后，其人数相对较少，不具有统计意义。因此，图 7(b)仅给出总价在 1000 以内所

对应的支持率，可以看出其总体趋势为总价越大，支持率越低。文本信息很难进行数值化的定量分析，因此，其影响直接通过训练的神经网络进行表示。

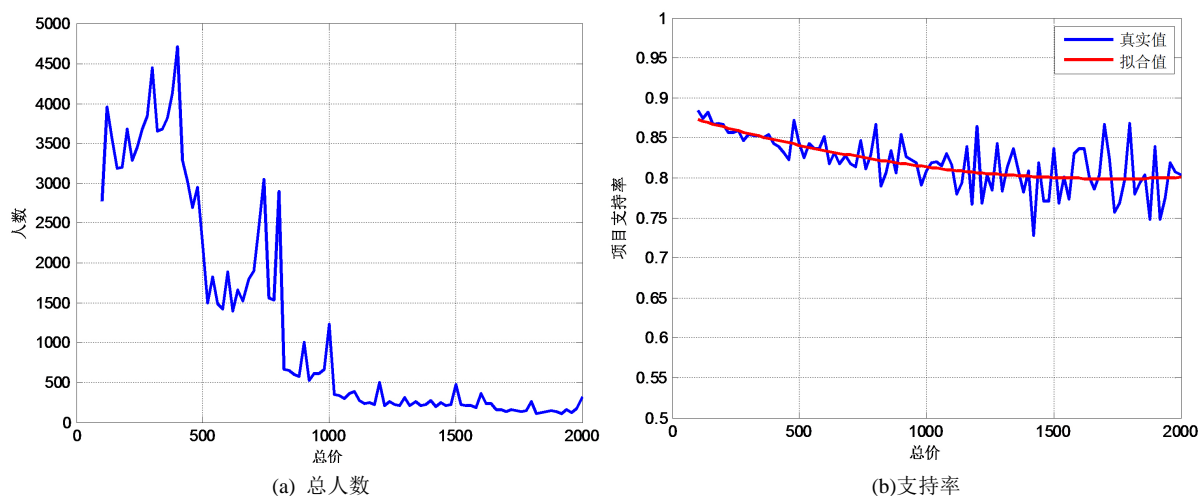


Figure 7. The impact of total price on project support rate
图 7. 总价对项目支持率的影响

4. 结论

本文将卷积神经网络和 BP 神经网络模型运用于多因素影响下的教育众筹成败预测中，利用卷积神经网络对纯文本信息进行网络训练；综合考虑了文本、总价、以前公布项目的教师人数等七种因素的影响，在原网络结构下引入 BP 神经网络，以评估其他因素的影响。利用已经训练好的网络，对 10,000 条测试样本输入网络并得到了分类结果，获得了 89.72% 的测试正确率。并且在 BP 神经网络众筹成败预测的基础上，通过各因素逐一缺失对项目支持率的影响，实现了各因素对众筹成败预测的量化。

参考文献

- [1] 黎明, 魏园园, 杨庆华. 浅议教育众筹模式发展[J]. 课程教育研究, 2016(22): 31-32.
- [2] 杨开城, 李通德, 惠治儒, 等. 在线教育众筹的技术机制与学习神经元[J]. 现代远程教育研究, 2017(1): 51-58.
- [3] 陈拓. 基于卷积神经网络的立体匹配技术研究[D]: [硕士学位论文]. 杭州: 浙江大学, 2017.
- [4] 朱灿. 基于 GA-BP 神经网络的奖励式众筹融资结果预测研究[D]: [硕士学位论文]. 上海: 上海师范大学, 2017.
- [5] 殷亚博, 王文忠, 等. 基于卷积神经网络和 KNN 的短文本分类算法研究[J]. 计算机工程, 2018, 44(7): 193-198.
- [6] Lecun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436. <https://doi.org/10.1038/nature14539>
- [7] 丁小刚. BP 神经网络与卷积神经网络在文字识别中的应用研究[D]: [硕士学位论文]. 武汉: 华中科技大学, 2014.
- [8] Schmidhuber, J. (2015) Deep Learning in Neural Networks. Elsevier Science Ltd., Amsterdam.
- [9] Kalchbrenner, N., Grefenstette, E. and Blunsom, P. (2014) A Convolutional Neural Network for Modelling Sentences.
- [10] 众筹网站[EB/OL]. <http://www.donorschoose.org>

知网检索的两种方式:

1. 打开知网首页: <http://cnki.net/>, 点击页面中“外文资源总库 CNKI SCHOLAR”, 跳转至: <http://scholar.cnki.net/new>, 搜索框内直接输入文章标题, 即可查询;
或点击“高级检索”, 下拉列表框选择: [ISSN], 输入期刊 ISSN: 2161-8801, 即可查询。
2. 通过知网首页 <http://cnki.net/>顶部“旧版入口”进入知网旧版: <http://www.cnki.net/old/>, 左侧选择“国际文献总库”进入, 搜索框直接输入文章标题, 即可查询。

投稿请点击: <http://www.hanspub.org/Submission.aspx>

期刊邮箱: csa@hanspub.org