Software Engineering and Applications 软件工程与应用, 2019, 8(6), 319-325
Published Online December 2019 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2019.86039

Hans汉斯

Prediction of Educational Crowds' Success or Failure Based on Convolutional Neural Network

Liping Luo, Jie Huang, Yage Zhang

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

Received: Nov. 8th, 2019; accepted: Nov. 26th, 2019; published: Dec. 3rd, 2019

Abstract

Educational crowdsourcing, to a certain extent, can optimize and test educational courses, effectively integrate social resources and educational resources, and alleviate the financial pressure for local governments. If crowdsourcing fails, it will cause huge time costs. Therefore, it is of great significance to predict the results of crowdsourcing projects [1]. Aiming at the problem of predicting the success or failure of educational crowdsourcing, the convolutional neural network model is applied to predict the success or failure of educational crowdsourcing. Word2vec, a language model of neural network, is used to train the word vectors of the text, and the trained word vectors are used to represent the text. The abstract features of the text are extracted by using convolutional neural network. On the basis of extracting abstract features, network training and prediction of text information are carried out, and 88.16% of the test accuracy is obtained.

Keywords

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

基于卷积神经网络的教育众筹成败预测

骆丽萍，黄　洁，张雅歌

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11月7日；录用日期：2019年11月26日；发布日期：2019年12月3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，

文章引用: 骆丽萍, 黄洁, 张雅歌. 基于卷积神经网络的教育众筹成败预测[J]. 软件工程与应用, 2019, 8(6): 319-325.
DOI: 10.12677/sea.2019.86039

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，将卷积神经网络模型运用于教育众筹成败预测中，通过采用神经网络语言模型word2vec对文本进行词向量的训练，并用训练好的词向量表示文本，使用卷积神经网络对文本进行抽象特征的提取，在提取出抽象特征的基础上，对纯文本信息进行网络训练和预测，获得了88.16%的测试正确率。

关键词

教育众筹，成败预测，卷积神经网络

1. 引言

众筹(Crowdfunding)是一种大众通过互联网进行沟通联系，并汇集资金支持由其他组织和个人发起的活动的群体性行为，也是通过互联网平台进行小额融资的新型融资模式[2]。学界对众筹的研究集中在2011年以后，主要来自金融和商业两个方面，在教育领域应用较少。教育众筹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，以此为学生提供更好的基础设施。跟传统办学模式相比，教育众筹直接面向广大网民募集资金，形式更加开放灵活，课程更加实用和新颖。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。

当前国内关于众筹的研究主要集中在对众筹模式的定性研究、众筹融资的影响因素、众筹投资者参与动机和行为等方面[3]，预测众筹融资结果的研究比较少。黄健青等人在逐步回归分析的基础上预测了项目融资的结果[4]；陈肖华等人将BP神经网络用于众筹项目融资结果的预测[5]。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，不仅能够更好的提取特征信息，同时还减少了网络的参数，便于模型的训练。

本文针对教育众筹的成败预测问题进行了深入研究，主要研究了教育众筹预测的基本原理、基于word2vec的特征提取、基于卷积神经网络的预测模型等几方面的问题，最后根据众筹网站的数据进行训练和预测分析，得到较理想的预测效果。因此本文可以为今后的教育众筹提供一定的理论建议，一定程度上帮助教育众筹者提高众筹的成功率。

2. 基于卷积神经网络的教育众筹成败预测

2.1. 教育众筹成败预测的基本原理

教育众筹的请求书上的主题信息大多以文本的形式呈现，需要大众做出接受或是不接受的判决，也就是说把文本分为两类：接受和不接受，因此教育众筹预测问题其实是个文本分类问题。文本分类系统的任务是：在给定的分类体系下，根据文本的内容自动地确定文本关联的类别。文本分类包括文本表示、训练过程和分类过程三部分内容。其中文本表示又可分为文本预处理、索引和统计、特征抽取等步骤；训练过

程就是用训练文本对事先选定的分类器进行训练和学习；最后依据训练好的准则对未知文本进行分类决策。

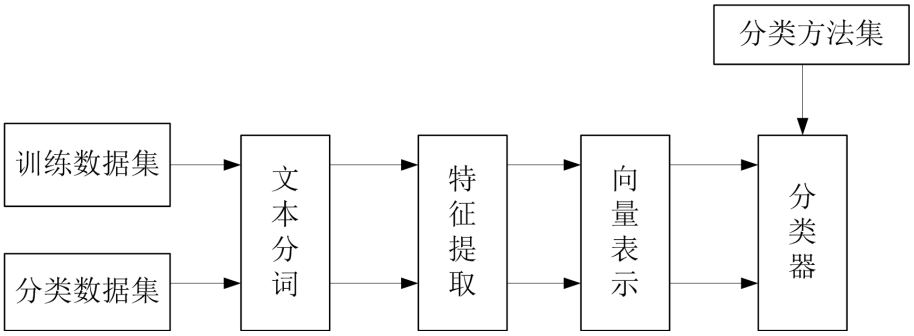


Figure 1. Text categorization process
图 1. 文本分类流程

一个典型的文本分类流程如图 1 所示。因此文本的表示和分类器的设计是本文要解决的两个关键技术。

2.2. 基于 word2vec 的词表征数值量化算法

目前，在文本信息处理问题上，文本的表示主要采用向量空间模型。特征项的选取和权重的计算是影响向量空间模型表述文本的重要因素，要想提高分类器的效率和分类准确性必须合理选择和提取文本的特征[6]。

word2vec 就是将词表征为实数值向量的一种高效的算法模型，其利用深度学习的思想，可以通过训练，把对文本内容的处理简化为 K 维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似[7]。其基本思想是通过训练将每个词映射成 K 维实数向量(K 一般为模型中的超参数)，语义相似度通过词与词之间的距离(比如 cosine 相似度、欧氏距离等)来衡量[8]。其采用一个三层的神经网络，输入层-隐含层-输出层。其核心是根据词频用 Huffman 编码，使得所有词频相似的词隐藏层激活的内容基本一致，出现频率越高的词语，他们激活的隐藏层数目越少，这样可以降低计算的复杂度。而 Word2vec 大受欢迎的一个原因正是其高效性，一个优化的单机版本一天可训练上千亿词[9]。

Word2vec 实际上是两种不同的方法：Continuous Bag of Words (CBOW)和 Skip-gram [10]。CBOW 的目标是根据上下文来预测当前词语的概率。Skip-gram 则根据当前词语来预测上下文的概率(如图 2 所示)。这两种方法都以人工神经网络作为它们的分类算法。起初，每个单词都是一个随机 N 维向量。经过训练之后，该算法利用 CBOW 或者 Skip-gram 的方法获得了每个单词的最优向量[10]。

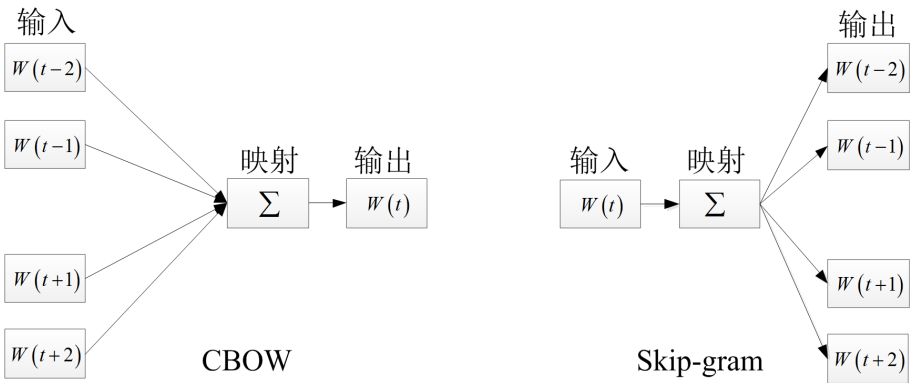


Figure 2. CBOW and Skip-gram method schematic diagram
图 2. CBOW 和 Skip-gram 方法示意图

2.3. 卷积神经网络分类器设计

卷积神经网络(Convolutional Neural Network, CNN)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。如图 3 所示，CNN 受到 Hubel 和 Wiesel 对猫大脑皮层的研究启发，用卷积层与池化层代替了传统神经元。实验证明，卷积与池化的交替结构对文本特征的学习与抽象有十分优良的效果。

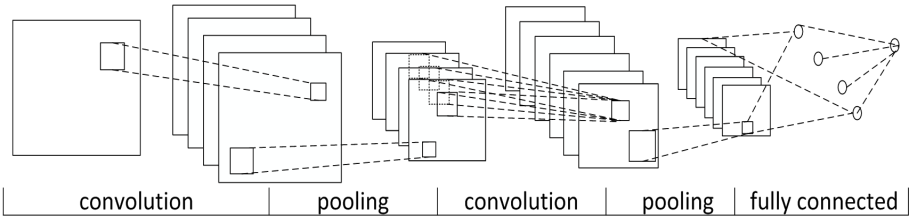


Figure 3. Structural sketch of convolutional neural network
图 3. 卷积神经网络结构示意图

1) 卷积层

卷积层本质上是一组滤波器，每个滤波器又称为卷积核。当文本输入网络时，单个卷积核通过固定步长下的不断滑动，与文本中的不同部分分别卷积，输出相应的特征二维矩阵。在神经网络中，滤波器的尺寸往往只覆盖部分特征，再通过滑动感知整个文本。滤波器尺寸的减少大大降低了神经网络的参数数量，提高了网络训练与预测的效率。

卷积层中的滤波器组权重不随输入的变化而变化，因此在卷积过程中，卷积核往往对某种特定特征输出较强，对其他特征响应较弱。因此每个滤波器都可以认为是一个特征提取方法，当文本局部输入符合滤波器条件时，输出的值越大，反之则输出值越小。

需要注意的是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[10]。

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

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[11]，本文使用的是 ReLU 函数。

ReLU 函数：

$$f(x) = \max(0, x)$$

卷积层的输入输出特征平面尺寸一般满足

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

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。

2) 池化层

池化本质是文本特征的一种聚合操作，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[11]，如图 4 所示，它是卷积神经网络另一种降维的手段。即使文本中目标特征有一个较小的平移或缩放，经过池化操作依旧能够得到和未变化前相同的池化特征。池化单元使卷积神经网络具有了一定的旋转、平移、伸缩不变性。

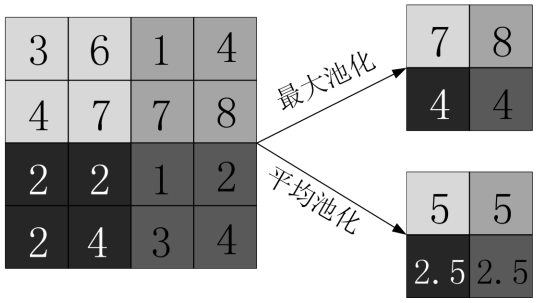


Figure 4. Maximum and average pooling diagrams
图 4. 最大池化和平均池化示意图

3) 全连接层

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，其缺点在于丢弃了文本原本的空间结构。因此 CNN 中，输入文本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。

本文算例 CNN 网络结构如表 1 所示：

Table 1. Network model structure table
表 1. 网络模型结构表

Input (200 × 128)		
Conv1D (64,128,3)	Conv1D (64,128,4)	Conv1D (64,128,5)
Relu	Relu	Relu
Maxpooling (3)	Maxpooling (4)	Maxpooling (5)
Reshape	Reshape	Reshape
	Merge	
	Dense (128)	
	Relu	
	Dense (2)	
	Softmax	

表中 Conv1D (a,b,c)表示卷积层，a 为卷积核个数，b、c 为感知野尺寸；Relu 表示激活函数，Maxpooling(d)表示最大池化层，d 为池化窗尺寸；Reshape 为变形层，主要用来将多维输入进行一维展开；Merge 为融合层，将多个输入融合；Dense(e)为全连接层，e 为神经元个数；网络输出最终通过 Softmax 函数。

$$S_i = \frac{e^i}{\sum_j e^j} \quad \forall i \in 1,2,\ldots,C$$

其中，网络输出为 i，输出个数为 C。

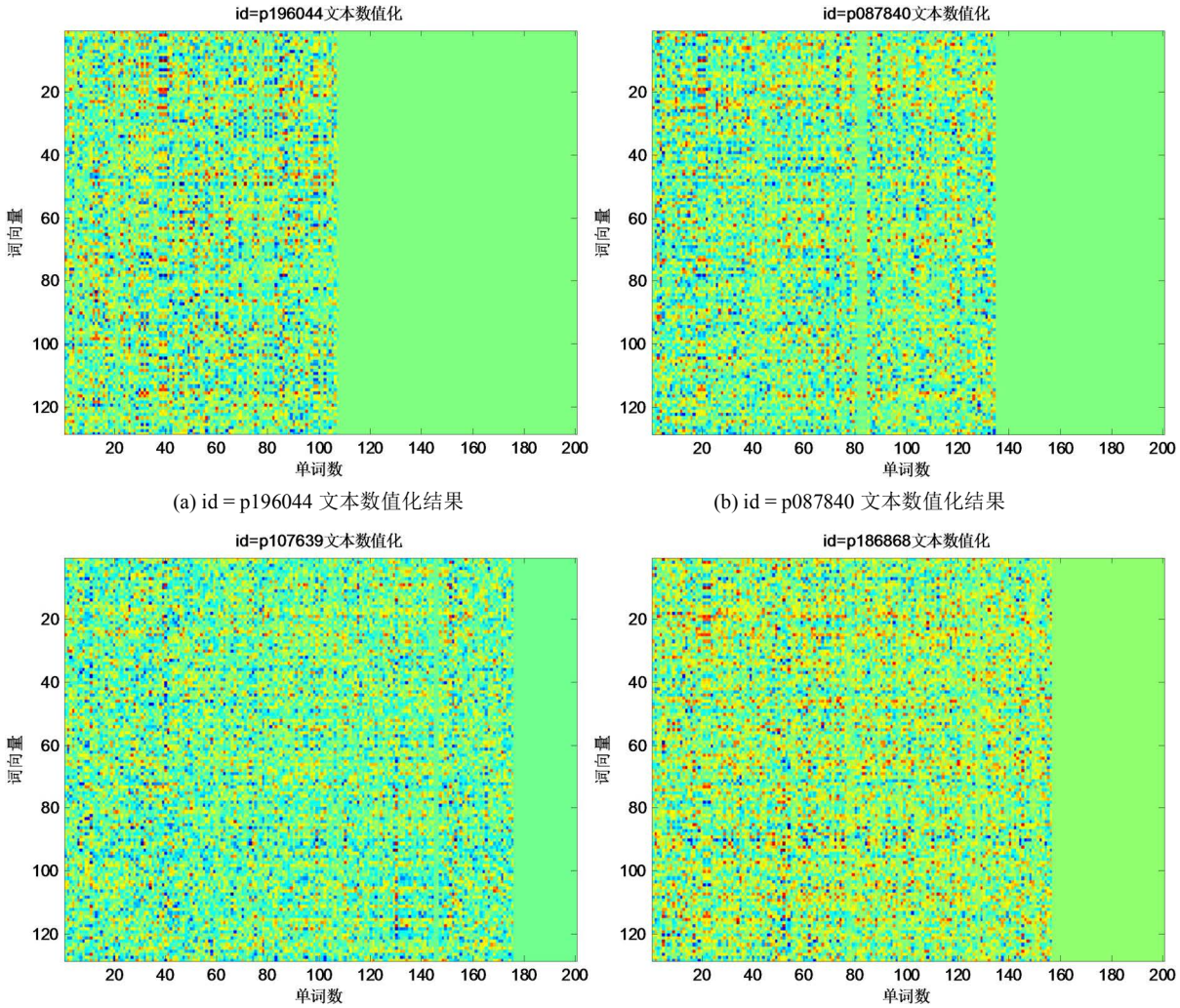
2.4. 实验结果与分析

本文根据众筹网站 https://www.donorschoose.org/的数据进行仿真分析[12]，该网站是全美国任意地区的教育众筹申请书公开众筹的平台之一。文章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 120,597 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 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。

2.4.1. 文本数值化处理结果

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 × 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，这里选取所有文本中最大的词语匹配数，这里选取 M = 200。

其中四个文本的数值化结果如图 5 所示：



(a) id = p196044 文本数值化结果　(b) id = p087840 文本数值化结果

(c) id = p107639 文本数值化结果　(d) id = p186868 文本数值化结果

Figure 5. Quantitative characteristic matrix of four texts
图 5. 四个文本的数值量化特征矩阵

2.4.2. 众筹成败预测结果

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 × 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 2 所示。

Table 2. Comparison of network performance under different perceptions
表 2. 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%

3. 结论

本文将卷积神经网络模型运用于教育众筹成败预测中，提出基于 word2vec 和卷积神经网络文本分类算法，完成了教育众筹的预测。该算法既有传统分类算法的简单高效性，又利用了 word2vec 将词表征为实数值向量的高效性，利用卷积神经网络对文本进行模型的构建、训练和分类。最后利用已经训练好的网络，对测试样本输入网络并得到了分类结果，获得了 88.16%的测试正确率，并比较了不同卷积和尺寸下的测试正确率。

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] 黄晓凤. 众筹项目融资成功的影响因素及预测模型研究[D]: [硕士学位论文]. 北京: 对外经济贸易大学, 2017.

[3] 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.

[4] 黄健青, 黄晓凤, 殷国鹏, 等. 众筹项目融资成功的影响因素及预测模型研究[J]. 企业管理与项目管理, 2017(7): 91-99.

[5] 陈肖华, 李元亨. 基于 BP 神经网络的众筹项目融资结果预测研究[J]. 科技创业月刊, 2017, 30(23): 31-33.

[6] 朱云霞. 结合聚类思想神经网络文本分类技术研究[J]. 计算机应用研究, 2012, 29(1): 155-156.

[7] 词表征数值量化. https://radimrehurek.com/gensim/models/word2vec.html

[8] 张轼坤, 沈峰, 高列宁, 周云康. 基于词向量的国际业务实时推理模型[J]. 信息技术与网络安全, 2019, 38(5): 85-89.

[9] 努比亚技术有限公司. 中文专利全文数据库[P]. 中国专利, 201611228051.5. 2016-05-31.

[10] 王磊, 等. 基于 Spark 的海量文本评论情感分析[J]. 苏州科技大学学报(自然科学版), 2018(1): 71-75.

[11] 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.

[12] 众筹网站. http://www.donorschoose.org