doi: 10.3969/j.issn.0490-6756.2019.01.011

基于 BP 神经网络的网络小说排行预测

龙 彬[1,2], 胡思才[1,3], 郭峻铭[1], 李旭伟[1]
(1. 四川大学计算机学院, 成都 610065; 2. 中国人民解放军 78179 部队, 成都 611130;
3. 中国人民解放军 61920 部队, 成都 610505)

摘 要: 近年来随着"IP"热潮兴起,网络文学市场发展迅速,逐渐成为文化娱乐行业投资热点.本文将机器学习方法引入到小说排行预测方面,通过网络爬虫获取网络小说信息并提取了影响排行的特征,提出了基于 BP 神经网络模型进行小说排行预测.针对训练数据的不均衡,本文采用 ROC 和 AUC 作为预测评价指标;实验结果表明,基于 BP 神经网络的网络小说排行预测的准确率较高,相比传统的文学定性分析方法,机器学习预测方法可解释性和应用性更高.

关键词: "IP"热潮; 小说排行预测; BP 神经网络; 网络爬虫; ROC 曲线; AUC 值
中图分类号: TP391 **文献标识码**: A **文章编号**: 0490-6756(2019)01-0050-07

Prediction of online novel rankings based on BP neural network

LONG Bin[1,2], *HU Si-Cai*[1,3], *GUO Jun-Ming*[1], *LI Xu-Wei*[1]
(1. College of Computer Science, Sichuan University, Chengdu 610065, China;
2. Unit 78179, PLA, Chengdu 611130, China; 3. Unit 61920, PLA, Chengdu 610505, China)

Abstract: With the rise of the "IP" boom in recent years, the online literature market is developing rapidly and has gradually become a hotspot for investment in the cultural and entertainment industry. This paper introduces the machine learning method to the novel ranking prediction, the characteristics of the influence rankings are extracted from the network novel (also called online novel) information collected by the web crawler, a BP neural network model is developed to predict the range of the online novel rankings. As the training dada is unbalanced, ROC and AUC are selected as the evaluation indicators of prediction. The experimental results show that the accuracy of online novel ranking prediction based on BP neural network is more accurate. Compared with traditional literary qualitative analysis method, machine learning approach is interpretable and applicable in the online novel ranking prediction.

Keywords: "IP" concept; the ranking of the novel predicts; BP neural network; web crawler; ROC curve; AUC value

1 引 言

近年来,"IP"(Intellectual Property,知识产权)概念逐渐成为热门,所谓"IP"主要是围绕拥有大量粉丝基数的文学作品相关版权改编,已变成文化产业中的重要操作模式,成为资本投资行业热捧的类型.拥有几亿读者的网络文学作品就是"IP"热潮的新宠儿,热映电视剧《余罪》、《琅琊榜》、《择天记》、《诛仙青云志》和高票房电影《九层妖塔》、《捉妖记》、《寻龙诀》等都是网络小说改编而来.2016 年数据表

收稿日期: 2018-05-05
基金项目: 国家自然科学基金(61173099)
作者简介: 龙彬(1985-),男,四川成都人,硕士生,研究方向为计算金融. E-mail: 280828129@qq.com
通讯作者: 李旭伟. E-mail: lixuwei@scu.edu.cn

明，我国 7.31 亿网民中，网络文学用户已达 3.33 亿，占网民总数的 45.6%，如图 1 所示. 文学网页日均浏览量超过 15 亿次，仅阅文集团一家网络小说存量达千万部[1]我国网络文学市场产值破 5000 亿，网络文学已成为我国数字出版产业重要组成部分，研究网络小说排行预测对网络文学市场投资有重要意义.

图 1　2012～2016 网络文学用户规模及使用率

Fig. 1　2012～2016 network literature user scale and usage rate

目前，我国的网络小说排行预测相关研究大部分是从文学角度定性分析，如吴琼以读者关注度为标准分析网络小说的市场潜力[2]，苏芯等人用分层回归方法研究了影响网络小说排行成功的因素[3]，姜岚等人从文学评价和批评的角度分析小说排行榜因素[4]，周志雄利用小说内容和类型解释网络小说排行[5]等等，这些研究缺少以大量数据为基础的定量分析，可解释性和应用性较弱. 利用机器学习方法预测排行可较有效避免这些问题，能更客观的区分出有投资潜力的网络小说.

网络小说虽然发展迅猛，但出现至今只有 20 年，时间相对较短，各网络小说网站提供的信息参差不齐，可供采集的各种数据还较为匮乏. 在数据多样性不够的条件下，回归预测准确度受影响较大，而分类预测的结果受影响较小. 分类预测中，输入直接使用采集的数据，数据预处理更简单，算法运行更快，因此分类预测比回归预测更适合用于小说排行预测.

本文结合我国网络小说的实际情况，提出一种基于 BP 神经网络的网络小说排行预测分类模型，该模型从多角度考虑到小说的排行影响因素，重点预测排名靠前的网络小说，发掘有潜力的原创网络文学作品，为网络文学投资提供参考.

2　BP 人工神经网络

BP 神经网络(下文简称 BP 网络)是一种使用 BP 算法训练的多层前馈神经网络，由一系列神经元组成[6]，包含输入层、隐含层和输出层，上下网络层神经元之间进行全连接，同层神经元之间无连接，如图 2 所示. 其中 F 是隐含层的线性或者非线性激活函数，由事前不描述的输入输出模式映射关系确定. 该网络使用 BP 算法原理：先是数据流前向传播，从输入层经过隐含层，到达输出层，计算出误差；再是反向传播，从输出层到隐含层，再到输入层，反复调节网络的参数直至最优[7]. BP 网络可以任意准确率逼近任何非线性映射得到预测结果，现已成功应用于电影票房预测[8]、地震预测[9,10]、邮件分类[11]、天气预报[12]以及图像分类[13]等方面.

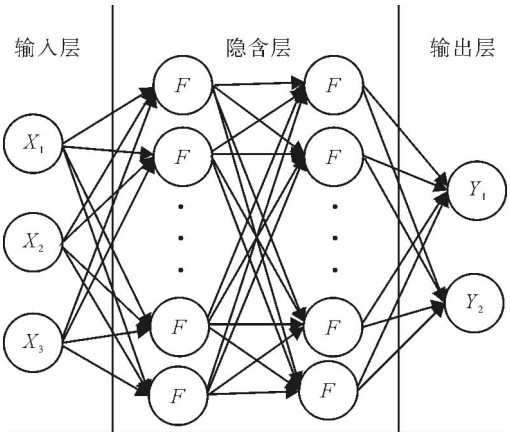

图 2　BP 网络结构

Fig. 2　BP network structure

3　BP 人工神经网络

本文的数据集是使用网络爬虫技术采集的，为匹配 BP 网络数据输入格式，需经过预处理后用于模型构建，构建流程如图 3 所示.

3.1　数据获取

为保证特征数据的代表性和丰富性，经过多方对比筛选，本文选择从国内最大文学阅读与写作平台、最大的原创文学门户网站起点中文网爬取数据. 2016 年我国发行的由网络文学改编的票房最高的 20 部电影里有 13 部，收视率最高的 20 部电视连续剧中有 15 部出自该网站网络小说改编. 该网站收录了 2002 年至今所有网络小说 1000 万余部，作品信息较为齐全，并按照总推荐票进行了小说名次排行，可直接利用排行划分分类标签，获取的小说数据如表 1 所示.

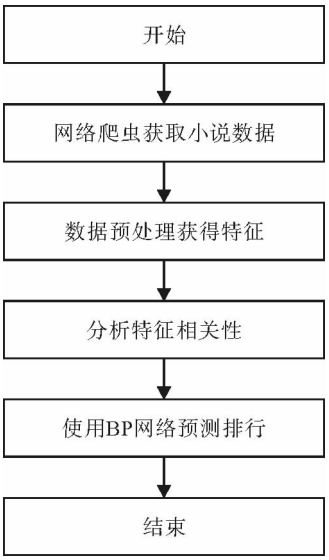

图 3 BP 网络小说预测流程
Fig. 3 BP network novel prediction flow

表 1 小说数据样例

Tab. 1 Novel data sample

书名	圣墟
作者	辰东
状态	玄幻
种类	连载
总字数	359.92 万
总点击	2717.52 万
总推荐	2056.82 万
评分	8.9
评分人数	5697
作者等级	白金
作品总数	6
总创作字数	2315.2 万
总创作天数	4017
作品荣誉	13.96

3.2 特征选择

相对于传统小说,内容质量对网络小说流行度影响并不显著,口碑、品牌信任、网站推荐、读者推荐、作者信任度和点击量、作者与读者的互动交流情况等要素对小说的排行产生的影响较显著[14]. 分析获取的小说数据,对比各个因素得到 9 组特征,分别是小说类型、小说评分、小说荣誉、小说字数、作者作品总数、作者等级、点击推荐比率、作者日写作效率、小说评价人数,为避免数字大小差异对模型训练带来干扰,所有特征值进行归一化处理.

小说类型是内容的高度概括,不同类型的小说会吸引不同的读者,本文按照起点中文网的方法分为 14 个网络小说类型:玄幻、奇幻、武侠、仙侠、都市、历史、军事、科幻、灵异、游戏、体育、二次元、短篇和女生作品,表达如下.

小说类型=类型小说总数/全站小说总数.

小说评分能反映小说的部分阅读者对小说质量和内容的评价,但因为读者个人爱好的差异性,评分并不能完全客观的体现小说质量[15,16],为了解决这一问题,将评分人数也考虑为排行影响因素.

评分数=小说评分/10;

评分人数=实际评价人数/10000,单位为万人.

作者是影响网络小说排行的最重要因素,起点中文网经过多年的实践已经形成了一套完整的作者分级制度,按照价值创作能力把作者影响力由小到大分为作家 Lv_1 ~ Lv_5,大神、白金 7 个等级,本文依次用 7 个数字表示作者的等级,作者等级表达如下式.

等级=当年收入指数×(1+5% * 写作年限)+影响力指数+基础分 10 分/10.

其中,影响力指数=当年月票总数/100;当年收入指数=作者当年的收入/1000.

作者等级不能完全体现作者对网络小说排行的影响,将作者作品总数和日创作效率两个因素也纳入特征范围,表达为

作者作品=作者作品数/10;

日创作效率=创作总字数/创作总天数,单位为万字.

爬取的小说荣誉是一段规范格式的文本,体现了小说受读者喜爱程度、获得的月票红包、达成的月票排行榜等成就,表达为

小说荣誉=荣誉动态字节长度/300.

小说字数指的是每本网络小说的字数,通常为 100 万字到 300 万字甚至更长,我们使用一个数字变量来表示小说的总字数,单位为亿字.

网络小说总点击表示其获得阅读次数,为排除恶意刷点击的可能,引入推荐票机制,起点中文网的每个注册用户每天只有 2 张推荐票,刷推荐票的可能较小,所以本文采取点击推荐比率作为阅读量特征.

点击推荐比=(小说总点击量/小说推荐总票数)/100

3.3　数据预处理

根据影响因素分析，进行数据预处理得到特征变量表如表 2 所示.

表 2　特征样例

Tab. 2　Feature sample

特征	小说 1	小说 2	小说 3
小说类型	0.305215	0.067328	0.171785
小说评分	0.86	0.82	0.84
作者等级	0.7	0.4	0.5
小说荣誉	0.1304	0.12	0.496
作品总数	1	0.7	0.9
日创作效率	0.748848	0.796195	0.500869
点推比	0.5274577	0.5501593	0.1342282
点评人数	0.3296	0.0244	0.06
小说字数	0.66901	0.19959	0.24554
标签	有潜力	潜力较小	有潜力

计算特征变量相系数矩阵，若相关系数绝对值较大(大于 0.7)，就可能存在多重共线性问题[17]，说明选择的特征不合理，反之，则说明特征选择较合理.

3.4　模型构建

本文选择 BP 网络模型构建一个二分类模型，步骤如下.

步骤 1　特征输入：将每个样本的 9 个特征 X 输入到网络的输入层神经元中，利用逻辑激活函数 g 将输入转化为一个 0 到 1 之间的概率值 $h_\theta(x_i)$，其中 g 为 Sigmod 函数，如下式.

$$g(x)=\frac{1}{1+e^{-x}} \tag{1}$$

$$h_\theta(x_i)=\frac{1}{1+e^{-(\theta_0+\theta_1x_1+\cdots+\theta_9x_9)}} \tag{2}$$

步骤 2　计算误差：前向传播过程中，将得到的输入 $h_\theta(x_i)$ 到误差函数中，计算误差 $J(\theta)$，表示为

$$J(\theta)=\frac{1}{m}\sum_{k=1}^{m}(y_i\log(h_\theta(x_i))+(1-y_i)\log(1-x_i)) \tag{3}$$

其中，y_i 为真实值，$h_\theta(x_i)$ 为预测值. 当 $y_i=1$ 时，$h_\theta(x_i)$ 接近 1，$\log(h_\theta(x_i))$ 就接近于 0，表示误差接近 0.

步骤 3　学习参数：计算误差 $J(\theta)$ 关于参数 θ 的偏导数 δ_i，求出每个样本的 δ_i 并取平均值，用所有特征的平均值调整原来的参数，δ_i 公式表示为

$$\delta_i=(y_i-h_\theta(x_i))h_\theta(x_i)(1-h_\theta(x_i))x_i \tag{4}$$

步骤 4　参数收敛：反向传播过程中，计算第三层的参数 δ^3，第二层的参数 δ^2，第一层是输入层无法求关于参数 θ 的偏导数，通过梯度下降迭代法学习到误差最小时的收敛参数，其中

$$\delta^3=h_\theta(X)-y \tag{5}$$

$$\delta^2=(\theta^2)^\theta\delta^3g'(z^{(2)}) \tag{6}$$

最后，将最终预测到的类别概率输出.

4　实验与分析

4.1　实验数据

4.1.1　*数据预处理*　本文使用数据为起点中文网总推荐排行前 3300 部作品信息，取 3.2 节所示的 9 种特征，计算出特征相关系数[18]矩阵热力图. 该图形为对称结构，横轴和纵轴 9 个颜色方块子图代表 9 种特征，对角线上的相关系数为 1. 子图颜色越接近黄色(1)，特征相关性越强且为正相关，子图颜色接近另一个极端深蓝色(−1)，特征之间相关性也很强且为负相关，0 处颜色表示特征之间不相关，右边颜色条表示颜色所对应的相关系数大小. 可以看出本文 9 种特征相关系数绝对值均在 0.7 以下，特征选择较为合理，如图 4 所示.

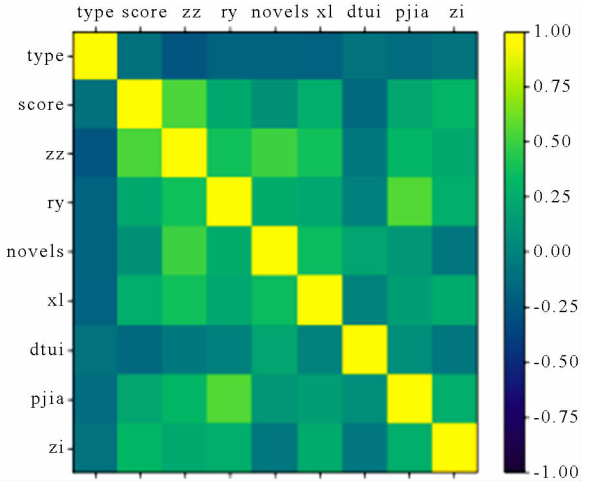

图 4　相关系数矩阵热力图

Fig. 4　Correlation coefficient matrix

4.1.2　*参数优化*　实验选择 python 语言实现一个 3 层 BP 网络模型，输入预处理后的数据，得到误差 $J(\theta)$，计算出各个特征的平均值 δ_1，如表 3 所示.

随机初始化网络参数：学习率=0.1，训练循环次数=10000，阈值=0.00001，随着训练次数增加，误差逐渐达到最小，参数收敛，如图 5 所示.

表 3 特征元素平均值

Tab. 3 Average of feature elements

第一本小说特征的值	0.612286614378
小说类型的偏导数平均值	−0.19986206
小说评分的偏导数平均值	−0.24050269
小说荣誉的偏导数平均值	−0.13073367
小说字数的偏导数平均值	0.00979895
作者作品总数的偏导数平均值	−0.140079
作者等级的偏导数平均值	−0.01762427
点击推荐率的偏导数平均值	−0.509761−2
日写作效率的偏导数平均值	0.02935095
小说评价人数的偏导数平均值	−0.00657909
以上参数计算出的分类准确值	0.95823215153

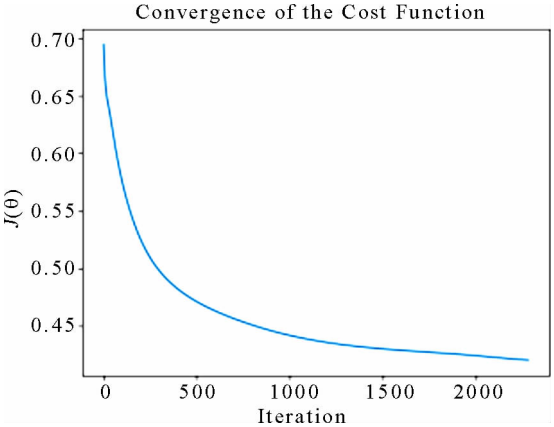

图 5 误差收敛

Fig. 5 Error convergence

4.1.3 实验结果 本实验数据集为 3300 部小说，取 2800 部小说作为训练集，500 部小说作为测试集. 输出层分为有潜力和潜力较小两类，起点中文网排行榜前 810 本小说总推荐票大于 30 万张，视为有投资潜力，占全部小说的 24.5%，剩下的小说投资潜力较小，占全部小说的 75.5%. 所以实验结果输出比例是非均衡的，为了更好的表达实验结果，采用 ROC 曲线和 AUC 值表示预测准确率. AUC(Area under Curve)表示 Roc 曲线下的面积，介于 0.1 和 1 之间，AUC 数值可以直观的评价分类器的好坏，值越接近于 1 越好[19].

在最优参数条件下，隐含层神经元数为 10 时多次实验，其中实验结果较好的 6 次数据如表 4 所示.

表 4 较好 AUC 值

Tab. 4 Better AUC value

实验次数	1	2	3	4	5	6
AUC 值	0.91	0.95	0.92	0.93	0.96	0.90

实验结果表明，可取得 AUC 值最大为 0.958232151153，其 ROC 曲线如图 6 所示. 其中，虚线表示随机分类 AUC=0.5 时的 ROC 曲线，阶梯形状实线为本实验 AUC 值对应的 ROC 曲线.

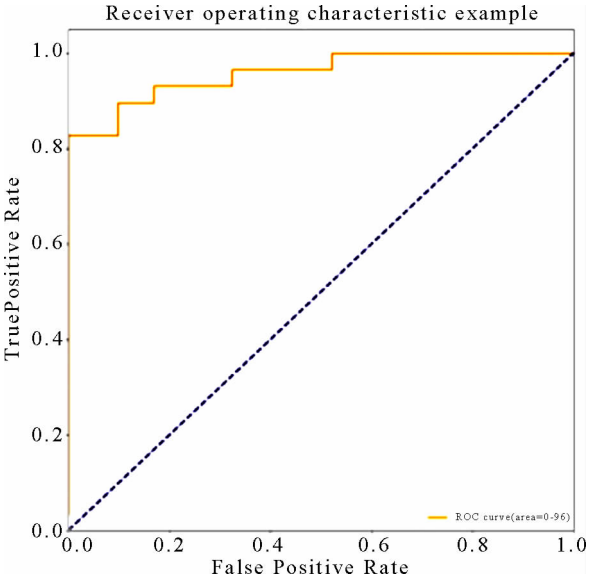

图 6 ROC 曲线

Fig. 6 ROC curve

4.2 实验比较

为了测试 BP 网络在小说预测方面的性能，在实验中将 BP 网络与 K-means 算法、朴素朴素贝叶斯模型、决策树模型进行预测准确率比较，选取多次实验中预测准确率较好的 6 次结果如表 5 和图 7 所示.

从预测准确率实验可以看出，同样实验环境和数据集条件下，K-means 算法预测准确率较差，朴素贝叶斯模型和决策树模型预测准确率一般，BP 网络的预测结果最好，准确率达到 96%，四种模型最好的预测准确率比较如表 6 所示.

表 5 模型准确率对比

Tab. 5 Model accuracy comparison

次数\准确率	1	2	3	4	5	6
K-means	0.90	0.89	0.88	0.88	0.87	0.89
朴素贝叶斯	0.89	0.92	0.91	0.90	0.92	0.91
决策树	0.91	0.92	0.93	0.93	0.94	0.92
BP 网络	0.93	0.92	0.94	0.93	0.96	0.95

表 6 最好准确率对比

Tab. 6 Best accuracy rate comparison

模型名称	BP 网络	决策树	朴素贝叶斯	K-means
准确率	0.958	0.940	0.924	0.908

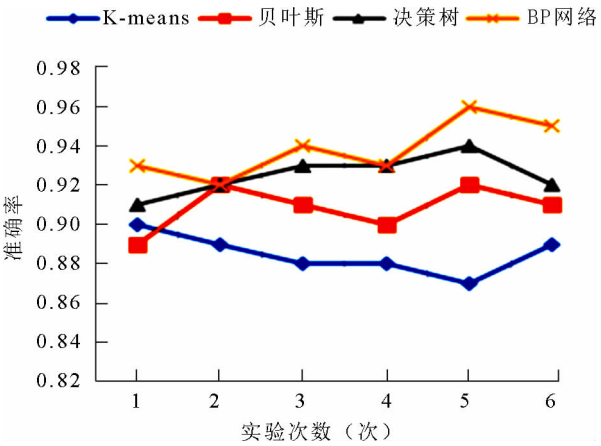

图 7　模型准确率对比

Fig. 7　Model accuracy comparison

在最佳准确率条件下，使用同一数据集在相同计算平台上多次执行实验程序，求程序执行时间平均值，对比四种模型时间复杂度，程序执行时间如表 7 所示.

表 7　执行时间对比

Tab. 7　Execution time comparison

模型名称	BP 网络	决策树	朴素贝叶斯	K-means
平均时间(s)	11.13	10.49	13.31	9.26

实验结果可见四种模型的程序执行时间都在 10 s 左右，其中 K-means 算法时间复杂度低，朴素贝叶斯模型执行时间高，BP 网络和决策树模型基本相同.

通过实验发现，BP 网络和朴素贝叶斯模型形成最优参数网络模型后，输入新数据不需再训练模型，决策树模型和 K-means 算法输入新数据都需要重新训练，表明 BP 网络和朴素贝叶斯模型在扩展性更好.

4.3　实验结果分析

通过预测准确率、时间复杂度、扩展性三方面的对比实验，可以看出本文采用的 BP 神经网络在预测网络小说排行榜方面预测准确率、扩展性好于同等数据集和计算环境下的 K-means 算法、决策树模型和朴素贝叶斯模型，BP 网络以牺牲一定的计算时间为代价换取较高的预测准确率和扩展性. 相比于传统的专家从文学角度主观定性分析预测小说排行，本文提出的以机器学习算法为基础的小说排行预测方法更加科学合理.

5　结　论

我国网络文学市场迅猛增长，逐渐成为文化娱乐投资热点，本文通过网络爬虫手段，获取起点中文网已排行的网络小说数据，在不考虑小说内容的前提下，选取网络小说信息特征，使用 BP 网络对小说排行榜进行预测. 通过实验分析对比，可以较为明显的区分出具有投资潜力的网络小说.

参考文献：

[1] Ito E, Urakawa T, Flanagan B, *et al*. Keywords frequency trend analysis of online novels [J]. IIAI Los Alamitos: IEEE, 2013, 2013: 68.

[2] 吴琼. 网络小说及其读者关注度分析 [J]. 图书馆建设, 2012, 2012: 66.

[3] 苏芯，刘益，李雪. 影响网络小说流行度的要素研究——以起点中文网为例 [J]. 上海管理科学, 2015, 37: 23.

[4] 姜岚，毕光明. 文学评价与批评眼光——来自小说排行榜的启示 [J]. 小说评论, 2015, 30: 12.

[5] 周志雄. 兴盛的网络武侠玄幻小说 [J]. 小说评论, 2016, 31: 116.

[6] 左言言，张海峰，庄婷. 车内声品质主客观评价模型对比分析 [J]. 江苏大学学报：自然科学版, 2017, 38: 403.

[7] 文家富，郭伟. 基于知识融合的汽车覆盖件模具设计方法研究 [J]. 重庆邮电大学学报：自然科学版, 2018, 30: 423.

[8] Zhang Z, Li B, Deng Z, *et al*. Research on movie box office forecasting based on internet data [C]// Proceedings of the International Symposium on Computational Intelligence and Design. Hangzhou: IEEE, 2016.

[9] 蔡润，武震，云欢，等. 基于 BP 和 SOM 神经网络相结合的地震预测研究 [J]. 四川大学学报：自然科学版, 2018, 55: 307.

[10] Narayanakumar S, Raja K. A BP artificial neueal network model for earthquake magnitude prediction in Himalayas, India [J]. Circuits Syst, 2016, 7: 3456.

[11] Tuteja S K, Bogiri N. Email Spam filtering using BPNN classification algorithm [C]//ICACAOT. Pune, India: IEEE, 2017.

[12] Wahyuni I, Adam N R, Mahmudy W F, *et al*. Modeling backpropagation neural network for rainfall prediction in tengger east Java [C]//Proceedings of the International Conference on Sustainable Information Engineering and Technology. Malang, Indonesia: IEEE, 2018.

[13] Wu K, Zhong Y F, Wang X M, *et al*. A novel approach to subpixel land-cover change detection

based on a supervised back-propagation neural network for remotely sensed images with different resolutions [J]. IEEE Geosci Remote S, 2017, 99: 1750.

[14] 郑坚, 周尚波. 基于神经网络的电影票房预测建模 [J]. 计算机应用, 2014, 34: 742.

[15] 陈伟鹏, 柳戌昊, 唐宁九. 基于灰色关联度和 AHP 的用户满意度综合评价模型 [J]. 四川大学学报: 自然科学版, 2017, 54: 713.

[16] 陶一嘉, 曹静. 解读信息爆炸时代的电影评分信任危机——以豆瓣电影平台为例的改良性设计 [J]. 工业设计研究, 2017, 5: 132.

[17] Twa M D, Parthasarathy S, Raasch T W, *et al*. Decision tree classification of spatial data patterns from videokeratography using zernike polynomials [C]//Proceedings of the Siam International Conference on Data Mining, San Francisco. Ca, USA: DBLP, 2003.

[18] Mallik R K. The exponential correlation matrix: eigen-analysis and applications [J]. IEEE Trans Wirele Commun, 2018: 1.

[19] Bagheri A, Sofotasios P C, Tsiftsis T A, *et al*. Area under ROC curve of energy detection over generalized fading channels [C]//Proceedings of the International Symposium on Personal, Indoor, and Mobile Radio Communications. Boston, MA, USA: IEEE, 2015.

引用本文格式:

中 文: 龙彬, 胡思才, 郭峻铭, 等. 基于 BP 神经网络的网络小说排行预测 [J]. 四川大学学报: 自然科学版, 2019, 56: 50.

英 文: Long B, Hu S C, Guo J M, *et al*. Prediction of online novel rankings based on BP neural network [J]. J Sichuan Univ: Nat Sci Ed, 2019, 56: 50.