浙江理工大学学报(自然科学版),第35卷,第6期,2016年11月
Journal of Zhejiang Sci-Tech University (Natural Sciences)
Vol. 35, No. 6, Nov. 2016

DOI:10.3969/j.issn.1673-3851.2016.11.014

基于递归神经网络的广告点击率预估研究

陈巧红,孙超红,余仕敏,贾宇波
(浙江理工大学信息学院,杭州 310018)

摘　要:为提高广告点击率的预估准确率,从而提高在线广告的收益,对广告数据进行特征提取和特征降维,采用一种基于LSTM的改进的递归神经网络作为广告点击率预估模型。分别采用随机梯度下降法和交叉熵函数作为预估模型的优化算法和目标函数。实验表明,与逻辑回归、BP神经网络和递归神经网络相比,基于LSTM改进的递归神经网络模型,能有效提高广告点击率的预估准确率。该模型不仅有助于广告服务商制定合理的价格策略,也有助于广告主合理投放广告,实现广告产业链中各个角色的收益最大化。

关键词:广告点击率;递归神经网络;LSTM;随机梯度下降;交叉熵
中图分类号:TP181　　文献标志码:A　　文章编号:1673-3851(2016)06-0880-06　　引用页码:110602

0　引　言

在中国,2014年在线广告首次超过电视广告,市场规模超过1500亿人民币,达到1540亿元,同比增长40.0%[1],2014年整年的比2012的773亿人民币,将近翻了一番,并且2015年有望突破2000亿人民币。

广告点击率预估作为计算广告学的重要研究领域,提高广告点击率是增加在线广告收益的重要手段之一。广告点击率预估模型主要是通过对丰富的历史数据进行挖掘,使预估模型尽可能学习历史数据中大量非线性特征之间的复杂关系,提高广告点击率预估的准确性。提高了广告点击率预估的准确性,结合广告的位置、广告竞价机制等因素使在线广告投放得更加准确,从而提高真实的广告点击率,根据在线广告付费机制,大多数公司都是采用按点击付费(cost per click,CPC),广告被点击次数越多,收益越高[2]。

广告点击率预估流程一般可以分为:特征提取,模型搭建,模型训练和模型预估这4个步骤。Joachims[3]从web搜索引擎日志中挖掘广告点击数据,并采用支持向量机对点击率进行预估。Graepel等[4]提出一种基于在线贝叶斯概率回归模型(online bayesian probability regression,OBPR)的预估方法,但由于该模型是基于特定的广告特征,所以难以实现个性化推荐。Chapelle等[5]提出一种基于动态贝叶斯网络模型的广告点击率预估方法,该方法引入"满意度"的概念,通过利用"满意度"分别模拟登陆页面的相关性以及搜索结果页面可感知的相关性,以此预估广告点击率,但是贝叶斯模型必须知道先验概率且属性之间必须是相互独立的。Dave等[6]则提出一种基于梯度增强决策树(gradient boosting decision tree,GBDT)的预估方法,该方法考虑了广告数据之间的相似性特征。Richardson等[7]采用逻辑回归模型(logistic regression model)进行广告点击率预估,并利用L-BFGS(limited memory broyden fletcher goldfarb shanno)训练模型。Agrawal等[8]提出一种基于时空模型(spatio-temporal predicting models)的预估方法。为解决稀疏数据广告点击率预估的问题,Agarwal等[9]提出一种稀疏数据预存的层次结构。Zhang等[10]为预估广告点击率提出COEC(clicks over expected clicks,COEC)模型,该模型将实际点击率与期望点击率之间的比值作为目标函数,该方法具有排序标准化的优点。Cheng等[11]则采用最大熵算法对广告点击率进行预估。而Zhang等[12]

收稿日期:2016-04-08
作者简介:陈巧红(1978-),女,浙江临海人,副教授,博士,主要从事计算机辅助设计及机器学习技术方面的研究。

提出一种基于递归神经网络(recurrent neural networks,RNN)的广告点击率预估方法,并利用反向传播算法(back propagation through time,BPTT)训练模型,实验表明该方法的预估准确率比普通的神经网络和逻辑回归模型高。但是,递归神经网络算法在使用梯度下降算法的时候会造成梯度消失或梯度爆发的问题,为了解决此类问题,本文采用基于长短期记忆(long short term memory,LSTM)改进的递归神经网络,利用 LSTM 特殊的结构,来避免在学习层次增加的情况下,梯度消失或者爆发的问题,从而提高模型的准确性。

本文的广告数据来自 Avazu 公司所提供的数据,针对大量丰富的广告日志数据,采用基于 LSTM 改进的递归神经网络模型去预估广告点击率。递归神经网络的隐藏层采用三层全连接结构,使得模型得以训练充分。实验证明基于 LSTM 改进的递归神经网络比逻辑回归模型、BP 神经网络(back propagation neural network,BPNN)和一般的递归神经网络的广告点击率效果越准确。

1　基于 LSTM 改进的递归神经网络模型

1.1　模型定义

LSTM 递归神经网络,该算法使用 LSTM 结构替换了一般的递归神经网络的隐藏层节点,LSTM 结构增加了输入门、输出门、遗忘门和一个内部单元(Cell)。其中一般的递归神经网络结构如图 1 所示。

输入门(Input Gate):表示是否允许输入层的信息进入到该隐藏层节点。门开的则允许输入层输出信号进入,门关则不允许信号进入,记为 ι。输出门(Output Gate):表示是否将当前节点的输出值输出给下一层。门开的则允许该隐藏层节点的信号输出,门关则不允许信号输出,记为 ω。忘记门(Forget Gate):表示是否保留当前隐藏层节点存储的历史信息。门开的则保留当前隐藏层节点存储的历史信息,门关则不保留当前隐藏层节点存储的历史信息,记为 φ。s_c^t 表示 t 时刻存储的信息值,模型的输入层和输出层与 RNN 模型一致。

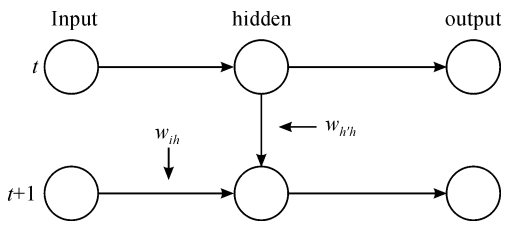

图 1　一般的递归神经网络

本文将隐藏层节点换成如图 2 所示的结构。

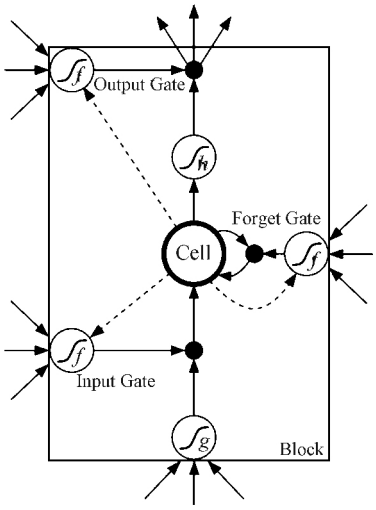

图 2　长短期记忆结构

基于长短期记忆(LSTM)改进的递归神经网络模型结构如图 3 所示。

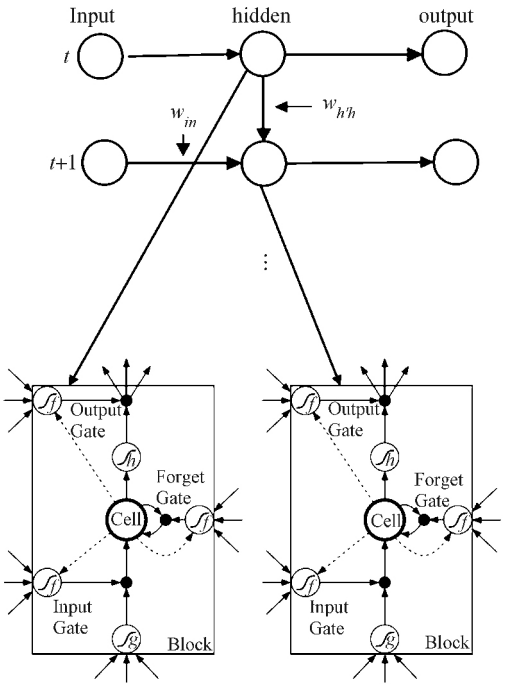

图 3　基于长短期记忆改进的递归神经网络模型结构

1.2　模型训练

与一般的递归神经网络的两个值来源作为输入不同,Input Gate 的输入由 3 个值构成。包括输入层节点的输出向量,前一个隐藏层的 cell 的输出向量,前一个时刻 cell 的保留信息。用 a_ι^t 表示表示 t 时刻输入门的输入向量,则通过输入门的激活函数得到 t 时刻该门的输出向量如式(1)所示:

$$b_\iota^t = f(a_\iota^t) \tag{1}$$

Forget Gate 的输入也同样由 3 个输入向量构

成，与输入门的来源是一样的。用 a_φ^t 表示 t 时刻遗忘门的输入向量，则通过遗忘门的激活函数得到 t 时刻该门的输出向量如式(2)所示：

$$b_\varphi^t = f(a_\varphi^t) \tag{2}$$

Cells单元从图3可知，它的输入由两部分组成，一个是输入层的输入向量，一个是前一个隐藏层的output门的输出。用 a_c^t 表示表示 t 时刻Cells单元的输入向量。

根据Forget门判断是否保留过去的信息值，如式(3)所示：

$$s_c^t = b_\varphi^t s_c^{t-1} + b_i^t g(a_c^t) \tag{3}$$

Output Gate的输入由3个部分组成，输入层的输出向量，前一个隐藏层的cell的输出向量和当前时刻的cell单元保留的信息，用 a_ω^t 表示表示 t 时刻输出门的输入向量，则通过输出门单元的激活函数得到 t 时刻该门的输出向量如式(4)所示：

$$b_\omega^t = f(a_\omega^t) \tag{4}$$

于是得到Cell单元的输出向量如式(5)所示：

$$b_c^t = b_\omega^t h(s_c^t) \tag{5}$$

其中：h 是激活函数。

Cell单元的输出向量，即隐藏层的输出向量，作为输出层的输入向量，如式(6)所示：

$$a_k^t = \sum_{c=1}^{H} w_{ck} b_c^t \tag{6}$$

最终从输出层输出的结果向量如式(7)所示：

$$b_k^t = f(a_k^t) \tag{7}$$

根据按时间的反向传播算法BPTT可得，t 时刻 i 节点到 j 节点的权值更新如式(8)所示：

$$w_{ij} = w_{ij} - \eta \delta_j^t b_j^t \tag{8}$$

其中：δ_j^t 表示 t 时刻 j 节点的残差值。

由于基于LSTM改进的递归神经网络模型中的状态是通过累加的方式计算的，其导数也是累加形式，避免了传统递归神经网络中由于导数逼近0而造成的梯度消失问题。

1.3 损失函数和评价函数

对数损失函数(log*loss*，见Cheese[13])跟AUC不同的是，AUC侧重广告点击率预估的排序，log*loss*侧重广告点击率预估的准确性，当点击率全部提高一定比例时，AUC因为只注重排序，所以不会有什么变化，而log*loss*值会引起变化。log*loss*是反映经过模型预估的点击率和真实点击率的拟合程度，值越小，广告点击率的预估结果越准确。本文采用scikit-learn里的log*loss*，核心如式(9)、式(10)所示：

$$\log loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}) \tag{9}$$

$$\log loss = -\frac{1}{N}\sum_{i=1}^{N}(y_i\log(p_i) + (1-y_i)\log(1-p_i)) \tag{10}$$

式(10)中的 y_i 表示表示第 i 个真实点击值，p_i 是通过模型预估的对于的点击值。

2 实验结果

2.1 数据分析

本文数据采用avazu公司提供的广告数据，为了方便对结果进行验证，这里仅采用训练集，训练集样本大小为40428967条。每条样本包括24个特征(其中包括15个显性特征和9个隐藏加密特征)，将训练集样本分成四份，将三份作为训练数据，一份为测试数据。采用多次划分，为了保证数据样本的可信度，分别进行训练。从表1可得，测试数据集和训练数据集的真实点击率是类似的，不会对模型的预测造成影响。

表1 测试数据集和训练数据集

Data1	广告的展现次数	点击数	真实的点击率/%
训练数据集	30000000	5102362	17.0
测试数据集	10428967	1762704	16.9

2.2 特征处理

通过对广告数据特征的分析可知，分析结果表明，在24个特征中其中Device_id和Device_ip的特征数过多，会导致很多长尾特征，如表2所示。

表2 Device_id和Device_ip的特征数

特征名	特征数
Device_id	2686408
Device_ip	6729486

由于以device开头的特征数过多所以产生了很多长尾的特征，为了使模型能更加稳定得学习，适当过滤掉一些长尾特征值。根据统计结果如表3所示。过滤掉device_ip出现频次低于10次的特征样本和device_id出现频次低于10的特征样本，最终得到训练样本数为23548762条。

表3 特征device_ip和device_id频次大于5和10的统计结果

特征名	频次大于5次	频次大于10次	原始的特征维度
device_ip	998149	462718	6729486
device_id	351216	87981	2686408

根据以 device 开头的特征,拼接字符,将 device_ip、device_id、device_model 和 C14 4 个特征拼接字符,然后再通过哈希映射将拼接的字符截取前 8 个字符,结果如表 4,记为 user_id,作为隐式用户属性。原始样本就有 24 个特征变成了 25 个特征,将特征 C15 和特征 C16 拼接成新的特征,记为 banner_size,也进行哈希编码。删除 C15 和 C16 特征,得到样本特征数变为 24 条。

将所得到特征进行归一化处理,将特征值映射到[0,1]之间,至此对特征处理完毕。

2.3　实验结果

对于逻辑回归模型,设置学习率为 0.0003,采用随机值初始化逻辑回归模型,优化函数采用 Adagrad。

对于 BP 神经网络模型,采用基本的三层结构,隐藏层节点根据经验所得,设置为 5 个节点。

对于 RNN 和基于 LSTM 的 RNN 模型,这里均采用 keras 框架,优化函数采用随机梯度下降(stochastic gradient descent,SGD)算法,目标函数都采用交叉熵函数,两个模型的特征处理也是采取同样的处理方法。

实验将训练数据分十等份进行训练,每次迭代一共包括 3 千万样本,每次迭代完计算 logloss 值。以上 4 个模型的迭代次数都一样,分别是 10、20、30、40、50、60。

不同模型 log*loss* 值随着迭代次数的不同,得到不同的值。表 4 列出了逻辑回归(logistic regression,LR)、BPNN、RNN 和基于 LSTM 的 RNN 模型在不同迭代次数时的 log*loss* 值。

表 4　实验结果

迭代次数	10	20	30	40	50	60
LR	0.39606	0.391579	0.389702	0.388364	0.389496	0.391109
BPNN	0.467845	0.464784	0.462847	0.461937	0.461315	0.461674
RNN	0.410213	0.4084561	0.402215	0.394515	0.390126	0.386266
LSTM	0.401235	0.395651	0.385263	0.383213	0.383756	0.383456

从表 4 中可以看出,对于 LR 模型随着迭代次数的增加,log*loss* 一开始在减少,当到达第 40 次迭代的时候就减少到最低了,log*loss* 值为 0.388364,第 50 次迭代后 log*loss* 值不减反增,可以看出模型在第四迭代时已经充分学习了。对于 BP 神经网络曲线一直处于下降状态,慢慢趋于平稳,但是比逻辑回归模型的 log*loss* 值高。对于 RNN 模型,随着迭代次数的增加,log*loss* 值处于下降状态,至 60 次时达到最小值 0.386266。对于基于 LSTM 的 RNN 模型,随着迭代次数的增加,log*loss* 值一开始处于下降状态,在 40 次时达到最低点 0.383213,这个值低于以上 3 个模型的最低值,如图 4 所示,表明这种方法预估的点击率比其它方法更准确。基于 LSTM 改进的递归神经网络采用特殊的 LSTM 单元结构,克服了一般的递归神经网络随着学习的深入,而造成梯度丢失的问题。

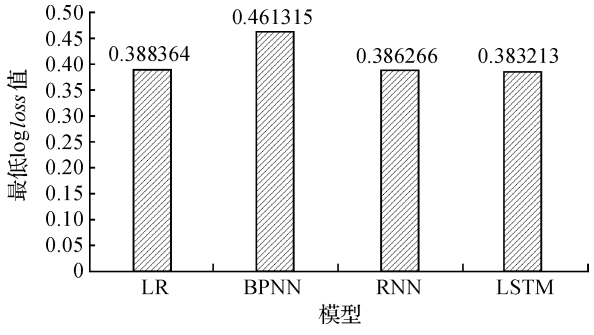

图 4　各个模型的最低 log*loss* 值对比

递归神经网络模型比逻辑回归模型和 BP 神经网络模型有更好的广告点击率预估的准确性,线性学习模型寻找一个线、面或者高维空间对一个数据特征趋势的无限逼近。线性模型本身就存在对非线性特征学习的不充分问题,无法完全体现众多特征之间的关系,而且随着迭代次数和学习时间的增加,容易造成过度拟合的问题,导致学习能力反而下降的情况。BP 神经网络模型本质上就是梯度下降算法,而且 BP 神经网络模型无法记忆信息,存在局部极小值的问题,会导致模型训练提早结束,BP 神经网络模型对初始状态设置较敏感,会出现同样条件的训练最后得到不同结果的情况,BP 神经网络采用的是误差修正法,所以收敛和训练的速度都非常慢。递归神经网络引进了时刻,相比隐藏层节点的输入只来自上一层节点的输出,递归神经网络模型的隐藏层节点不仅来自上一层节点的输出还来自上一时刻的隐藏层节点的输出,递归神经网络模型可以学习更加复杂特征之间的关系。

基于 LSTM 改进的递归神经网络模型采用特殊结构替代了普通的曲线神经元激活函数,当上一时刻的误差反向传递来的时候,可通过其中的记忆单元记下来,从而可以很好地记录历史信息,尽量防止梯度消失的问题。实验结果证明了在广告点击率

预估上，基于LSTM改进的递归神经网络模型比其他的模型有更好的预估效果。

3 结 语

本文采用基于LSTM改进的递归神经网络模型去对历史广告数据进行预估，通过实验对比逻辑回归模型、BP神经网络模型和标准的递归神经网络模型，实验结果证明本文所采用的基于LSTM改进的递归神经网络模型在预估广告点击率方面比对比模型的准确率要好，克服了一般递归神经网络随着学习层次的加深，造成的梯度消失问题，进一步证明了本文所做工作的有效性。

基于神经网络的模型对输入特征都比较敏感，超高维的特征不仅影响模型训练，甚至有可能造成模型无法训练，所以如何更有效得对海量数据进行特征提取、特征选择和特征降维，成为未来的热门研究问题。此外，神经网络的优化算法和目标函数也有很多，本文递归神经网络采用了的优化算法是SGD算法，新的优化算法也在不断提出，例如自适应算法Adadelta算法[14]、Adagrad算法[15]等。目标函数本文采用交叉熵函数，常见的目标函数还包括均方差、绝对值均差和多分类逻辑回归。优化算法和目标函数也是未来一个值得研究的问题。

参考文献：

[1] 艾瑞咨询. 艾瑞咨询：2015年中国网络广告行业年度数据监测[EB/OL]. (2015-04-20)[2015-12-20]. http://www.iyunying.org/seo/sjfx/12636.html.

[2] 周傲英，周敏奇，宫学庆. 计算广告：以数据为核心的Web综合应用[J]. 计算机学报，2011，34(10)：1805-1819.

[3] JOACHIMS T. Optimizing search engines using clickthrough data[C]//Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining. ACM，2002：133-142.

[4] GRAEPEL T，CANDELA J Q，BORCHERT T，et al. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine [C]//Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010：13-20.

[5] CHAPELLE O，ZHANG Y. A dynamic bayesian network click model for web search ranking [C]// Proceedings of the 18th International Conference on World Wide Web. ACM，2009：1-10.

[6] DAVE K，VARMA V. Predicting the Click-Through Rate for Rare/New Ads [R]. Centre for Search and Information Extraction LabInternational Institute of Information Technology . Hyderabad，2010.

[7] RICHARDSON M，DOMINOWSKA E，RAGNO R. Predicting clicks：estimating the click-through rate for new ads [C]//Proceedings of the 16th International Conference on World Wide Web. ACM，2007：521-530.

[8] AGARWAL D，BRODER A Z，CHAKRABARTI D，et al. Estimating rates of rare events at multiple resolutions[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2007：16-25.

[9] AGARWAL D，AGRAWAL R，KHANNA R，et al. Estimating rates of rare events with multiple hierarchies through scalable log-linear models[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2010：213-222.

[10] ZHANG W V，JONES R. Comparing click logs and editorial labels for training query rewriting[C]//WWW 2007 Workshop on Query Log Analysis：Social And Technological Challenges. 2007.

[11] CHENG H，CANTú-PAZ E. Personalized click prediction in sponsored search[C]//Proceedings of the Third ACM International Conference on Web Search and Data Mining. ACM，2010：351-360.

[12] ZHANG Y，DAI H，XU C，et al. Sequential Click Prediction for Sponsored Search with Recurrent Neural Networks [C]//Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI，2014：1369-1375.

[13] CHEESE T. Logarithmic Loss[EB/OL]. (2016-02-25)[2016-02-26]. https://www.kaggle.com/wiki/LogarithmicLoss.

[14] ZEILER M D. Hierarchical convolutional deep learning in computer vision [D]. New York：New York University，2013.

[15] WAGER S，WANG S，LIANG P S. Dropout training as adaptive regularization [C]//Advances in Neural Information Processing Systems. 2013：351-359.

Research on Estimation of Ads Click Rate Based on Recurrent Neural Network

CHEN Qiaohong, *SUN Chaohong*, *YU Shimin*, *JIA Yubo*

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: In order to improve the estimation accuracy of ads click rate and thus improve the revenue of online advertising, feature extraction and dimension reduction of advertising data were implemented. Then, the improved recurrent neural network based on LSTM was used as the ads click rate estimation model. Meanwhile, stochastic gradient descent and cross entropy were used as optimization algorithm and objective function separately. Experiments show that compared with logistic regression, BP neural network and recurrent neural network, the improved recurrent neural network based on LSTM can effectively improve the estimation accuracy of ads click rate. It not only helps advertising service providers develop reasonable price strategies, but also helps advertisers advertise reasonably. As a result, the revenue maximization of each role in the advertising industry chain is realized.

Key words: ads click rate; recurrent neural network; LSTM; stochastic gradient descent; cross entropy

(责任编辑: 陈和榜)