基于卷积神经网络的互联网金融信用风险预测研究*

王重仁,韩冬梅
(上海财经大学 信息管理与工程学院,上海 200433)

摘　要:针对互联网金融行业的信用风险评估问题,提出了一种基于卷积神经网络的客户违约风险预测方法。首先将输入数据分为动态数据和静态数据,将动态数据和静态数据分别转换为矩阵和向量,然后利用改进的卷积神经网络来自动提取特征并进行分类,最后使用 ROC 曲线、AUC 值和 KS 值作为评价指标,将该方法与其他机器学习算法(Logistic 回归、随机森林)进行比较。实验结果表明,卷积神经网络模型对于信用风险的预测效果要优于对比模型。

关键词:信用风险评估;卷积神经网络;机器学习;深度学习
中图分类号:TP391　　**文献标识码**:A　　DOI: 10.19358/j.issn.1674-7720.2017.24.013
引用格式:王重仁,韩冬梅.基于卷积神经网络的互联网金融信用风险预测研究[J].微型机与应用,2017,36(24):44-46,50.

Prediction of credit riskin Internet financial industry based on convolutional neural network

Wang Chongren, Han Dongmei
(Department of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China)

Abstract: A method of customer default risk prediction based on convolutional neural network is proposed in the light of credit risk evaluation problem in the Internet financial industry. Firstly, the input data is divided into dynamic data and static data, and the dynamic data and static data are converted into matrix and vector. Then, an improved convolutional neural network is used to automatically extract features and classify. Finally, the ROC curve, AUC value and KS value are used as evaluation metrics, and the method is compared with other machine learning algorithms (Logistic Regression and Random Forest). The experimental results show that the convolutional neural network model overcomes the contrast model in the field of customer credit risk prediction.

Key words: credit risk evaluation; convolutional neural networks; machine learning; deep learning

0　引言

近年来,国内互联网金融行业呈爆炸式增长态势,随着行业的不断发展,如何有效评价借款人的信用风险已成为互联网金融行业能否可持续健康发展的关键环节之一,日益受到人们的重视[1]。

客户信用风险评估本质上是一个分类问题,即将客户分成违约和按时还款两类。客户信用风险预测模型的发展经历了三个阶段:定性分析、统计学方法和人工智能方法[2]。定性分析是最早用于信用评估的方法,其后统计学方法被逐渐引入到信用评估中。近年来,随着机器学习的发展,一些智能化方法被陆续应用到信用评估研究中。例如,MALEKIPIRBAZARI M 等[3]使用随机森林算法对国外网络借贷平台 Lending Club 借款人的风险进行预测。然而,这些传统机器学习方法预测效果的好坏非常依赖于人工设计的特征,而人工设计特征的方法往往无法考虑到所有特征,同时人工设计特征需要花费大量时间和人工成本[4]。

近年来,深度学习受到了越来越多学者的关注,卷积神经网络(Convolutional Neural Network, CNN)则是其中一种经典而广泛应用的网络结构。LECUN V 等人[5]在 1998 年提出了 LeNet-5,LeNet-5 成功应用到了手写字符识别领域。2012 年,KRIZHEVSKY A 等人提出的 AlexNet[6]在 ImageNet 图像分类竞赛中夺得了冠军,使得 CNN 成为了各界关注的焦点。在此之后,CNN 模型不断改进,比如 Google 的 GoogLeNet[7]等。CNN 能够从数据中自动学习特征,从而代替人工设计特征,且深层的结构使它具有很强的表达能力和学习能力。经过不断发展,CNN 逐渐从图像分类扩展到其他领域,比如:行人检测、自然语言处理、语音识别等。目前 CNN 的应用场景大部分都是非结构化数据分类问题,近年来,开始有研究尝试将 CNN 应用到结构化数据分类问题中,比如李思琴等[8]提出了基于 CNN 的搜索广告点击率预测的方法。本文研究所用数据来源于国内一家互联网金融平台——融 360,本文尝试使用卷积神经网络来进行互联网金融行业违约风险预测研究。

1　方法

1.1　数据编码

将输入数据分为两类,一类为静态数据,如描述用户

* 基金项目:上海财经大学研究生教育创新计划项目(2015111101)

基本属性的性别、职业等；另一类为动态数据，动态数据主要包括用户的历史行为数据，如用户的银行流水记录、用户浏览行为、信用卡账单记录。动态数据为时点数据，会随着时间的变化而改变。本文研究所用到的输入数据的变量如表 1 所示，数据的标签为用户是否违约，用户违约定义为逾期 30 天以上。

表 1　输入数据变量列表

变量类型	变量
用户的基本属性	性别
	职业
	教育程度
	婚姻状态
	户口类型
银行流水记录	银行流水时间
	收入金额
	支出金额
	工资收入金额
信用卡账单记录	账单时间
	上期账单金额
	上期还款金额
	信用卡额度
	本期账单金额
	本期账单最低还款额
	消费笔数
	预借现金额度
	还款状态
用户浏览行为	用户浏览时间
	浏览行为类型
	浏览行为数据
放款时间	放款时间

在本文中，将用户动态数据转换成矩阵，矩阵如图 1 所示，其中矩阵的行代表用户的行为数据，矩阵的列代表时间，时间基本单位为月或周。假设用户放款时间为时间点 t，用户数据的时间范围为放款前 m 个时间单位和放款后 n 个时间单位，因此构建矩阵时，矩阵各列以用户放款时间 t 为基准，按照时间的先后顺序进行排列。假设用户的行为数据种类数量为 p，那么矩阵共有 p 行，$m+n+1$ 列，矩阵中的元素代表用户在某个时间点上的某一种行为的特征，一个矩阵代表了一个用户在不同时间点上的所有行为特征。

将银行流水记录和信用卡账单记录的时间基本单位设置为月，以月为单位进行汇总，将用户浏览行为的时间基本单位设置为周，以周为单位进行汇总，汇总时可选用的指标有合计、计数、平均等。因三种历史行为记录转换成的矩阵的大小不相同，所以将三个矩阵作为三个单独的数据源进行输入。

对于输入数据中静态数据，因数据不会随着时间的改变而改变，所以用向量的方式来进行编码，假设用户基本属性数据在数据处理后的种类数量为 q，则用户静态输入数据的大小是 $1\times q$。

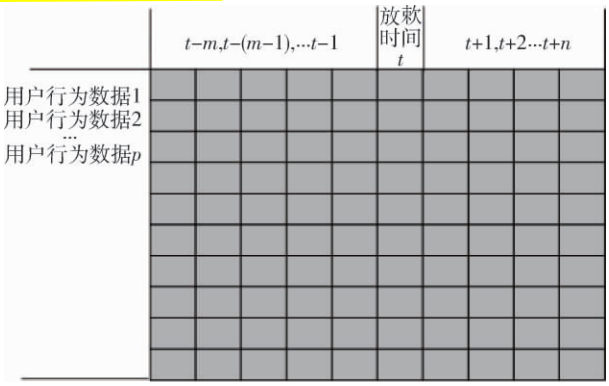

图 1　动态数据转换后的矩阵示意图

1.2　卷积神经网络结构

本文提出的改进卷积神经网络模型借鉴了经典的 LeNet-5 和 GoogLeNet 的结构，构建的卷积神经网络模型包含四个子卷积网络，每个子卷积网络都有单独的输入，四个子卷积网络最后在全连接层(Fully Connected Layer，FC) 进行融合，全连接层之后是 Softmax 输出层，CNN 结构如图 2 所示。

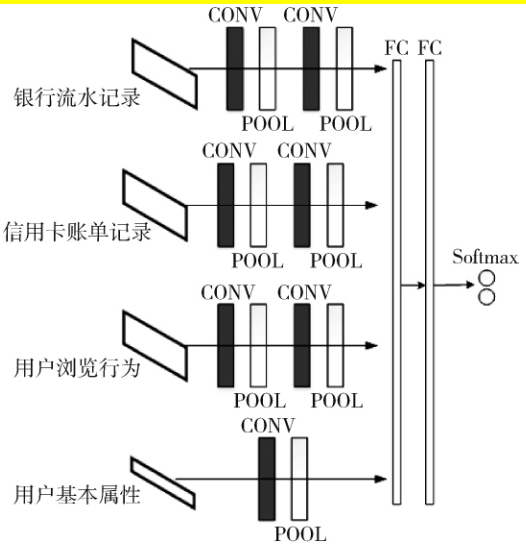

图 2　卷积神经网络结构

对于四个子卷积网络，与动态输入数据连接的子卷积网络采用两个卷积层(Convolutional Layer，CONV) 和两个池化层(Pooling Layer，POOL) 来自动提取特征，考虑到静态输入数据特征较少，与静态数据连接的子卷积网络只采用了一个卷积层和一个池化层。

卷积层由多个特征面(Feature Map) 组成，每个特征面由多个神经元组成，每一个神经元通过卷积核与上一层特征面的局部区域连接。卷积层利用局部连接和权值共享，减少网络自由参数个数，降低网络参数复杂度[9]。卷积层计算公式如下：

$$X^{(l)} = f(W^l \otimes X^{(l-1)} + b^{(l)}) \quad (1)$$

其中 $X^{(l)}$ 和 $X^{(l-1)}$ 代表层 l 层和 $l-1$ 层的神经元活性，W^l 代表卷积核，b 代表偏置。

对于与动态输入数据连接的三个子卷积网络，采用相同的参数，在第一个卷积层，选择了 64 个大小为 1×3 卷积核，目的是提取用户每一个行为在不同时间点上的特征。卷积层之后是池化层，池化层起到二次提取特征的作用，它的每个神经元对局部接受域进行池化操作。常用的池化方法有最大池化、随机池化和均值池化，这里选择最大池化法（取局部接受域中值最大的点）。在池化层之后连接第二个卷积层，选择了 128 个大小为 3×3 的卷积核，目的是进一步提取用户每一个行为指标在不同时间点上的特征，并且提取用户同一时间上不同行为的特征。在第二个卷积层后同样连接了一个池化层。对于静态输入数据采用了一个卷积层和一个池化层来提取特征，卷积层使用了 64 个大小为 1×3 的卷积核。

四个子卷积网络的输出全部在全连接层进行融合，第一个全连接层和第二个全连接层的维度分别是 512 和 256。最后，选择 Softmax 函数作为输出分类器。Softmax 函数估计输入 x 属于特定类别 $j \in k$ 的概率：

$$P(y = j \mid x) = \frac{\exp(x^{\mathrm{T}} W_j)}{\sum_{k=1}^{K} \exp(x^{\mathrm{T}} W_k)} \quad (2)$$

选择常用的修正线性单元（Rectified Linear Unit，ReLU）作为激励函数，ReLU 激励函数可以防止梯度消失和过拟合问题，ReLU 激励函数定义为：

$$f_{\mathrm{cov}}(x) = \max(0, x) \quad (3)$$

Dropout 是 CNN 中防止过拟合提高效果的一种有效手段，它是指在卷积神经网络的训练过程中，对于神经网络单元，按照一定的概率将其从网络中丢弃，本文在每个子卷积网络的最后一个池化层后面进行 Dropout(0.3) 操作。

为了证明 CNN 在用户信用风险预测问题上的优越性，选择了在信用风险预测领域常用的两种传统机器学习方法作为对比：Logistic 回归（Logistic Regression，LR）和随机森林（Random Forests，RF）。

1.3 评价指标

以 TP（True Positive）代表被模型预测为正的正样本，以 TN（True Negative）代表被模型预测为负的负样本，以 FP（False Positive）代表被模型预测为正的负样本，以 FN（False Negative）代表被模型预测为负的正样本。

ROC（Receiver Operating Characteristic）和 AUC（Area under Curve）指标是常用的评价指标。首先计算真正率（TPR）和假正率（FPR）的值，然后以 FPR 和 TPR 为坐标形成折线图，即 ROC 曲线。

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (4)$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} \quad (5)$$

ROC 曲线越靠近左上角，代表模型分类性能越好。AUC 是 ROC 曲线下方面积，AUC 越大，代表模型的准确性就越高。

KS（Kolmogorov-Smirnov）是信用风险评分领域常用的评价指标，首先将数据样本按照预测违约概率由低到高进行排序，然后计算每一个违约率下的累积 TPR 值和累积 FPR 值，最后求这两个值的差值的最大值，即为 KS 指标。KS 值越大代表模型对于违约客户和按时还款客户的分辨能力越强。

2 实验结果

2.1 实验设置

本文数据源共包含 50 000 个用户的数据。首先对数据进行预处理。将类别型变量，如性别，转换为 One-hot 编码，同时将连续型变量，如收入金额，进行归一化处理。将用户行为记录和用户基本属性分别转换成矩阵和向量作为 CNN 的输入。同时采用特征提取的方式，从用户行为记录中抽取特征作为传统算法的输入，特征值从用户行为记录中汇总得到，选用的汇总指标有合计、计数、平均等。为了更好地对模型进行评估，将数据划分为训练集、验证集和测试集。

2.2 结果分析

实验结果如表 2 和图 3 所示，表 2 中显示了 3 种模型实验结果的 AUC 值和 KS 值。从表中可以看到，本文构建的 CNN 模型实验结果的 AUC 值和 KS 值都远远高于传统方法。同时如图 3 所示，CNN 的 ROC 曲线始终处于最左上方。以上表明本文提出的 CNN 方法具有较好的信用风险预测效果。

表 2 模型运行结果

模型	AUC	KS
LR	0.62	0.17
RF	0.57	0.11
CNN	0.75	0.42

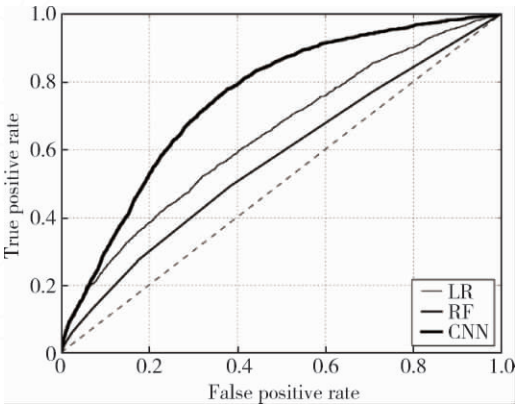

图 3 ROC 曲线

3 结论

本文针对互联网金融行业的用户信用风险评估问题，

（下转第 50 页）

分辨率低的情况。但是,分辨率为 128 ×128 时的识别率低于分辨率为 64 ×64 的情形。另外,经测算,基于主成分特征的方法正确识别率不低于类似的识别方法,但减少了计算复杂量。在今后的研究中,还需要针对更复杂的环境,克服困难采集大量的交通标志图像,进行训练和测试,以进一步提高正确识别率。

参考文献

[1] 杨守建, 陈恳. 基于 Hopfield 神经网络的交通标志识别[J]. 计算机工程与科学,2011,33(8):132-137.

[2] 朱双东,陆晓峰. 道路交通标志识别的研究现状及展望[J]. 计算机工程与科学,2006,28(12): 50-52,102.

[3] ZAKLOUTA F, STANCIULESCU B. Real-time traffic-sign recognition using tree classifiers[J]. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(4):1507-1514.

[4] FISTREK T, LONCARIC S. Traffic sign detection and recognition using neural networks and histogram based selection of segmentation method[A]. IEEE Processing of ELMAR,2011: 51-54.

[5] 许少秋. 户外交通标志检测和形状识别[J]. 中国图象图形学报,2009,14(4):708-711.

[6] 罗晓萍,蒋加伏,唐贤瑛. 基于 SVM 和模糊免疫网络的交通标志图像识别[J]. 计算机工程与设计,2006,27(9):1542-1544.

[7] 朱正平,孙传庆,王秀丽,等. 基于外观特征与神经网络的交通标志识别[J]. 自动化与仪器仪表,2009(1): 60-63.

[8] 孙光民,王晶,于光宇,等. 自然背景中交通标志的检测与识别[J]. 北京工业大学学报,2010,6(10): 1337-1343,1395.

[9] BAE G Y, HA J M, JEON J Y, et al. LED traffic sign detection using rectangular hough transform[C]. IEEE Conference Publications, 2014 International Conference on Information Science and Applications(ICISA), 2014:1-4.

[10] PARK J G, KIM K J. A method for feature extraction of traffic sign detection and the system for real world scene[C]. 2012 IEEE International Conference on Emerging Signal Processing Applications(ESPA),2012: 13-16.

[11] 陈亦欣,叶锋,肖锋,等. 基于 HSV 空间和形状特征的交通标志检测识别研究[J]. 江汉大学学报(自然科学版),2016,44(2):119-125.

[12] 胡月志,李娜,胡钊政,等. 基于 ORB 全局特征与最近邻的交通标志快速识别算法[J]. 交通信息与安全,2016,34(1):23-29.

[13] 蒋先刚. 数字图像模式识别工程项目研究[M]. 成都: 西南交通大学出版社,2014.

(收稿日期: 2017-05-31)

作者简介:

邹柏贤(1966 -),男,博士,副教授,CCF 高级会员(E200025653S),主要研究方向: 机器学习、图像处理。

苗军(1970 -),男,博士,副教授,主要研究方向: 人工智能、神经网络、图像理解。

(上接第 46 页)

提出了一种基于卷积神经网络的客户违约风险预测模型。首先将输入数据分为动态数据和静态数据,将动态数据和静态数据分别转换为矩阵和向量,本文建立的卷积神经网络模型包含四个子卷积网络,最后使用 ROC、AUC 值和 KS 值作为评价指标,将该方法与其他传统机器学习算法(LR、RF)进行比较。实验结果表明,卷积神经网络模型的客户违约风险预测性能要优于其他模型,能对借款人的信用风险进行更准确的评估,同时,卷积神经网络模型能够从数据中自动学习特征,与人工设计特征相比,可以节约大量的时间,因此本文建立的模型在互联网金融行业的信用风险评估领域更具有优势。

参考文献

[1] 于晓虹,楼文高. 基于随机森林的 P2P 网贷信用风险评价、预警与实证研究[J]. 金融理论与实践,2016(2):53-58.

[2] REDMOND U, CUNNINGHAM P. A temporal network analysis reveals the unprofitability of arbitrage in the prosper marketplace[J]. Expert Systems with Applications, 2013, 40(9): 3715-3721.

[3] MALEKIPIRBAZARI M, AKSAKALLI V. Risk assessment in social lending via random forests[J]. Expert Systems with Applications, 2015, 42(10): 4621-4631.

[4] 操小文, 薄华. 基于卷积神经网络的手势识别研究[J]. 微型机与应用, 2016, 35(9): 55-57.

[5] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[6] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]. International Conference on Neural Information Processing Systems, 2012: 1097-1105.

[7] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]. Computer Vision and Pattern Recognition, 2015:1-9.

[8] 李思琴, 林磊, 孙承杰,等. 基于卷积神经网络的搜索广告点击率预测[J]. 智能计算机与应用, 2015(5): 22-25,28.

[9] 郑昌艳, 梅卫. 基于卷积神经网络的空中目标战术机动模式分类器设计[J]. 微型机与应用, 2015, 34(22): 50-52.

(收稿日期: 2017-05-25)

作者简介:

王重仁(1984 -),男,博士研究生,主要研究方向: 数据挖掘。

韩冬梅(1961 -),女,博士生导师,教授,主要研究方向: 经济分析与预测。