融合数据增广技术与机器学习算法的个人信用评分研究

陆健健[1,2],江开忠[2]

(1.上海工程技术大学 管理学院;2.上海工程技术大学 数理与统计学院,上海 201600)

摘　要:为了提高个人信用评分模型算法预测精准率,受视觉领域数据增广思路启发,提出融合数据增广技术与机器学习算法的个人信用评分模型。该模型首先对原始个人信用数据进行数据增广处理,然后基于机器学习分类算法训练一个二分类个人信用评分模型,最后基于公开个人信用数据集,分别建立未经过数据增广和经过数据增广处理后的个人信用评分模型。对比准确率、精确率、召回率、F1得分、AUC值和ROC曲线等6个性能评价指标,结果显示,相较于仅基于机器学习算法的个人信用评分模型,融合了数据增广技术与机器学习算法的个人信用评分模型使得分类性能得到了一定提升,分类准确率平均高出5%。

关键词:数据增广技术;机器学习算法;个人信用评分;分类性能评价指标

DOI:10.11907/rjdk.192197　　开放科学(资源服务)标识码(OSID):

中图分类号:TP306　　**文献标识码:**A　　**文章编号:**1672-7800(2020)008-0040-04

Research on Personal Credit Score of Fusion Data Augmentation Technology and Machine Learning Algorithm

LU Jian-jian[1,2], JIANG Kai-zhong[2]

(1. *School of Management*, *Shanghai University of Engineering Science*;
2. *College of Mathematics and Statistics*, *Shanghai University of Engineering Science*, *Shanghai* 201600, *China*)

Abstract: Inspired by data augmentation in computer vision, it is feasible to increase the number of training data and make the data set as diverse as possible so as to improve the accuracy of the model of personal credit scoring. After the data is augmented, the performance of the classification task can often be greatly improved. This paper firstly proposes a personal credit scoring model based on data augmentation algorithm. Based on the data augmentation of original personal credit data, a personal credit model is established based on supervised machine learning algorithm. In the empirical part, this paper builds a personal credit scoring model that has not undergone data augmentation and data augmentation processing based on public personal credit data sets. Six performance evaluation indicators, such as accuracy, accuracy, recall, F1 score, AUC value and ROC curve showed that the classification performance was improved more than 5% by the personal credit scoring model based on data augmentation technology.

Key Words: data augmentation; machine learning; credit scoring; classification performance evaluation metrics

0　引言

近年来,随着人工智能和机器学习算法的不断进步与发展,作为人工智能和机器学习领域的一个典型应用,个人信用评分模型取得了长足进步。个人信用评分问题实质上是模式识别领域的一个分类问题,识别个人信用特征并将个人判断划分为不违约和违约两类。具体做法是根据历史数据样本信息,从已知数据中识别违约及不违约者特征,从而总结出分类规则,构建分类算法模型,用于测量借款个人违约概率,为相关决策者或者决策机构提供决策依据[1]。本文将过去个人信用评分研究算法模型主要划分为三大类:传统数学建模方法[2-6];决策树、K近邻判别(KNN)、聚类、神经网络等单一机器学习算法[7-10];集成算

收稿日期:2019-11-12

基金项目:上海工程技术大学研究生创新项目(18-01114)

作者简介:陆健健(1993-),男,上海工程技术大学管理学院硕士研究生,研究方向为统计学习、机器学习;江开忠(1965-),男,博士,上海工程技术大学数理与统计学院副教授、硕士生导师,研究方向为统计分析。

法或者多算法融合[11-15]。面对众多评分模型,模型侧的更新换代已成为个人信用评分研究领域发展的瓶颈。

近些年,在计算机视觉、自然语言处理等领域出现的数据增广技术可以为个人信用评分研究提供新思路,成为提升个人信用评分模型预测精度的突破口。2018 年,卢海涛等[16]通过图像变换和合成技术建立满文古籍文档图像数据增广算法,解决训练数据不足问题,在构建的增广数据集上建立 Faster R-CNN 深度学习模型挖掘深层图像特征,实现满文文档图像印章检测方法,并对采集的真实满文文档复印件图像进行测试,印章检测精度可以达到 99.6%。同年,蒋梦莹等[17]提出优化分类的数据增广方法,通过对测试集所有类别进行分析,找到分类效果不好的单类进行数据扩增,改善模型因训练样本少、结构复杂引起分类效果差的现象,为数据增广方法提供了多种思路。2019 年,王钰清等[18]基于数据增广和卷积神经网络算法的地震随机减噪,对无噪地震数据添加不同方差的高斯噪声,数据增广后构成新的训练集,实现了对小样本 CNN 模型训练。

参考以上视觉领域图片处理的数据增广技术,本文对传统个人信用数据使用 SMOTE 算法进行增广。与传统信用评分模型相比,本文将数据增广思想运用于个人信用数据这类二维结构化数据集中,提出了一种融合数据增广技术与机器学习算法的个人信用评分模型。相比之前未经过数据增广的模型,该模型算法具有预测精准度高、鲁棒性好等特点。

1　相关技术原理

1.1　数据增广技术

数据增广技术是深度学习中的常用技巧,主要用于增加训练数据集数据量,让数据集尽可能多样化,使得训练的模型具有更强的泛化能力。在实际各项应用中,并非所有数据增广方式都适用于当前训练数据集,需要根据自己的数据集特征确定应该使用哪几种数据增广方式。目前,在视觉领域,数据增广主要包括:水平/垂直翻转、旋转、缩放、裁剪、剪切、平移、对比度、色彩抖动等方式;在自然语言处理领域,数据增广主要包括:同义词替换、随机插入、随机交换、随机删除等技术;而在二维结构化数据领域,目前尚未有学者提出统一数据增广技术,而仅仅在出现不平衡数据集时,有研究者提出了基于 SMOTE 算法、SMOTE 算法的以变种为代表的过采样技术,这种过采样技术实际上就是针对不平衡数据集中少数类数据的数据增广技术,如果将对象换作全体各类数据集,将全体数据集做过采样处理,则那些过采样技术就是本文所指的数据增广技术。

1.2　机器学习算法

常用的机器学习算法主要分为无监督学习和有监督学习。本文主要用到了有监督学习,有监督学习主要是指输入的样本数据有相应的标记类别。有监督学习算法可以从给定的训练数据集中学习出一个模型参数,当给定一个新的数据样本时,可以根据该模型参数预测一个相应类别的结果。有监督学习的训练集要求包含输入和输出,也可以说是特征属性和目标属性。监督学习包括回归预测问题和分类预测问题,通过已有的训练样本去训练得到一个误差最小的最优模型,再利用该最优模型对输入样本输出相应结果,最后对输出进行简单判断从而实现预测目的,也即对未知数据样本具有预测的能力。常见的有监督学习分类算法有 K 近邻、支持向量机、决策树、随机森林、梯度提升树、XGBoost 等。本文在实验部分将使用以上几种有监督学习的分类算法。

1.3　数据增广算法流程

本文数据增广技术原理与 SOMTE 算法思想一致,区别在于传统 SOMTE 算法只扩增少数类样本,使少数类样本数据与多数类样本达到平衡,而本文数据增广原理是指扩充所有样本数据,使得依据样本训练出来的模型达到精确度高、避免过拟合的效果。

(1)首先,对于数据集中每一行样本记录 X,以欧氏距离为标准计算它到它所属类别样本集 S 中所有样本的距离,取其中距离最近的 K 个样本记录,得到其 k 近邻。

(2)其次,根据样本数据集设置一个增广比例以确定增广倍率 N,对于每一个类样本记录 X,从其 k 近邻的样本中随机选择若干样本,记选择的近邻样本为 XN。

(3)最后,对于每一个随机选出的样本 XN,分别与原样本按照式(1)构建新的样本。

$$x_{new}=x+rand(0,1)\times(\hat{x}-x) \tag{1}$$

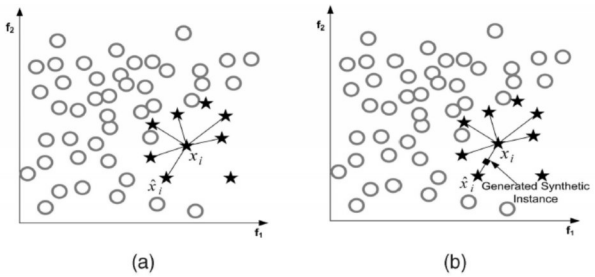

图 1　数据增广实例

2　实验与结果分析

2.1　数据集描述及预处理

为更好地验证经过数据增广的模型具有更高的准确率,本文选取两个公开 Benchmark 的数据集,它们均来源于加州大学 UCI 公开数据库,其中包括著名 German、Australian 两个信用数据集,它们都是关于银行信用卡个人用户业务信息的数据,如表 1 所示。

表 1　数据集基本信息

数据集	类别	数目	属性数目	数值属性数	类别属性数
German	正类	300	20	7	13
	负类	700			
Australian	正类	383	14	8	6
	负类	307			

德国数据集共有样本记录 1 000 条，其中正类 300，负类 700，属性数目共 20 个，其中数值型属性 7 个，类别属性 13 个。澳大利亚数据集共有样本记录 690 条，其中正类 383，负类 307，属性数目共 14 个，其中数值型属性 8 个，类别属性 6 个。这两个数据集的具体属性信息如表 2 和表 3 所示，其中澳大利亚数据集公开贡献者为了保护数据隐私，所有属性名和值都被替换成一些没有意义的变量。

在实际问题的数据集中经常会出现缺失值的情况，而缺失值往往也会导致模型的准确率不高，因此在训练原始数据集之前，需要对原始数据样本进行数据预处理。首先，对原始数据集中严重缺失数据的样本记录予以剔除，对部分缺失数值型样本采用均值填充方法，对分类型变量部分缺失数值的样本记录采用众数填充的方法；其次，对所有分类型变量的数据进行编码，本文采用的是 OneHot 编码；最后，对所有数值型数据进行规范化处理，本文对数据采取极差标准化，如式(2)所示，其中 X 代表某属性原始数据，$X_{\min}$ 代表某属性数据的最小值，$X_{\max}$ 代表某属性数据的最大值，X^* 代表标准化后某属性的数据。

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

在对原始数据进行填充和标准化过后，将数据拆分为训练集和测试集，本文拆分的比例为 7∶3。K 折交叉验证技术在模型训练和检验中也常被使用，本文使用了最常用的 10 折交叉验证。

2.2 性能评价指标

二分类模型评价准则中常用的混淆矩阵如表 2 所示，几个常见模型评价术语有：正类代表信用差的样本，负类代表信用好的样本数据。本文用 TP 表示实际为正类，预测也为正类的样本个数；用 FN 表示实际为正类，预测为负类的样本个数；用 FP 表示实际为负类，预测为正类的样本个数；用 TN 表示实际为负类，预测也为负类的样本个数。

表 2 二分类混淆矩阵

分类	预测正类	预测负类
实际正类	TP	FN
实际负类	FP	TN

(1)准确率(Accuracy)。在二分类算法模型评价指标体系中，准确率(Accuracy)是一个非常重要也很直观的指标，表示所有准确分类的样本数与所有样本数之比。通常而言，准确率越高，分类器越好。其数学表达如式(3)所示。

$$\text{准确率(Accuracy)} = (TP + TN)/(TP + FN + FP + TN). \tag{3}$$

(2)精确率(Precision)、召回率(Recall)和 F1 得分。准确率高并不能说明算法一定好。而且在实际应用中，并不总是关心预测准确率，在特定领域，人们往往更加关注该模型对某一特定类的判别能力。在信用领域，对于金融机构而言，它更在意的是信用差的人被判为信用好所带来的损失，也即它希望假负类的比率能够越低越好。因此在评价一个模型好坏时，往往还要引入精确率(Precision)、召回率(Recall)和 F1 这 3 个指标。精确率是指分类正确的正样本个数占分类器判别为正样本的样本个数比例，召回率是指分类正确的正样本个数占真正正样本个数比例。精准率与召回率是既矛盾又统一的两个指标，为调和这两个指标的矛盾，设计出了 F1 得分指标，F1 得分是指精确率与召回率的调和平均。其数学表达如式(4)—式(6)所示。

$$\text{精确率(Precision)} = TP/(TP + FP) \tag{4}$$

$$\text{召回率} = \text{(Recall)} = TP/(TP + FN) \tag{5}$$

$$F1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}. \tag{6}$$

(3)ROC 曲线及 AUC 值。除以上指标外，ROC 曲线与 AUC 值也常被用来评价模型。ROC 曲线又称真正率-伪正率图，其中横坐标用伪正率表示，纵坐标用真正率表示。虽然直接用 ROC 曲线在视觉上很直观，但无法去定量评价不同的分类模型，因此通常采用 ROC 曲线下方的面积作为评价指标，即 AUC 值。AUC 值在 0~1 之间，通常而言，AUC 值越大，分类模型的性能也就越好。真正率和伪正率如式(7)、式(8)所示。

$$\text{真正率} = TP/(TP + FN) \tag{7}$$

$$\text{伪正率} = FP/(FP + TN) \tag{8}$$

2.3 实验结果

为了验证经过数据增广后的算法模型具有更好的性能，本文对德国个人信用数据集建立逻辑回归、支持向量机、朴素贝叶斯、K 近邻、决策树、随机森林、极限梯度提升、梯度提升树等 8 对算法模型，结果如表 3 所示。

表 3 八对算法模型在德国信用数据集上性能对比

数据集	算法模型	准确率	精确率	召回率	F1 得分	AUC 值
German	lr	0.784	0.82	0.78	0.79	0.85
	lr_da	0.748	0.75	0.75	0.75	0.82
	sgdc	0.608	0.65	0.61	0.6	0.72
	sgdc_da	0.56	0.87	0.56	0.64	0.73
	svm	0.516	0.90	0.42	0.5	0.70
	svm_da	0.73	0.75	0.73	0.73	0.69
	mnb	0.676	0.7	0.68	0.69	0.72
	mnb_da	0.636	0.65	0.64	0.64	0.71
	knn	0.716	0.83	0.72	0.75	0.70
	knn_da	0.866	0.87	0.87	0.87	0.95
	dtc	0.74	0.74	0.74	0.74	0.70
	dtc_da	0.843	0.84	0.84	0.84	0.84
	rfc	0.764	0.85	0.76	0.79	0.81
	rfc_da	0.881	0.88	0.88	0.88	0.94
	xgboost	0.792	0.82	0.79	0.8	0.85
	xgboost_da	0.847	0.85	0.85	0.85	0.93
	gbdt	0.784	0.8	0.78	0.79	0.83
	gbdt_da	0.849	0.85	0.85	0.85	0.93

由表 3 可以看出，在德国信用数据集上，除回归(lr)、朴素贝叶斯(mnb)与数据增广技术融合后的模型较原模型性能低外，其它 6 个融合模型都比原模型性能好，特别是 k

近邻(knn)、决策树(dtc)、随机森林(rfc)、极限梯度提升(XGBoost)、梯度提升树(GBDT)等融合后的模型在所有性能指标上都比原模型要高出不少,准确率平均高出 6%左右。

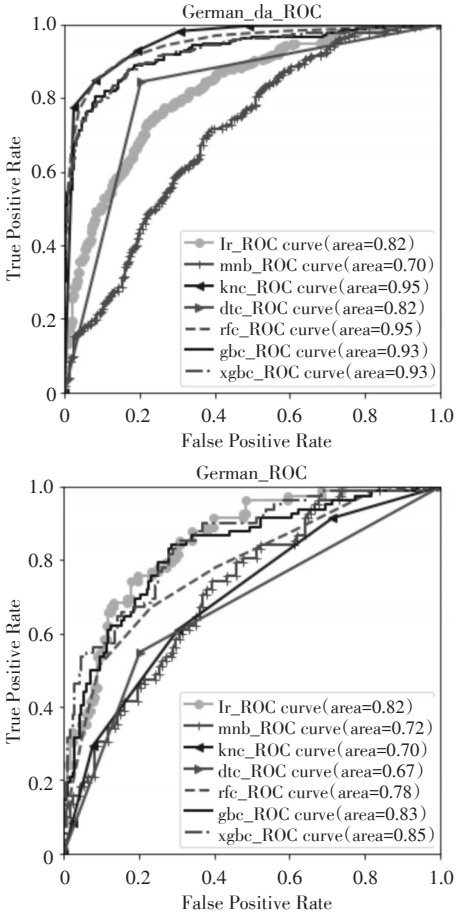

图 2　各算法在两个数据集上的 ROC 曲线及其 AUC 值

由图 2 可以看出,两个 ROC 曲线凸出,也即在德国数据集上,经过与数据增广技术融合的算法性能都得到了显著提升。

3　结语

随着机器学习深度学习算法的不断发展,个人信用评分也得到了巨大发展,但是面对众多算法模型,算法模型侧的升级换代已成为个人信用评分研究领域发展的瓶颈。本文参考视觉和自然语言处理领域的数据增广思想,提出了一种数据增广技术与算法相融合的思路。基于两个公开信用数据集,对比 8 组机器学习算法模型实验,结果显示,采用融合数据增广技术的算法显著提高了个人信用评分模型的预测准确率及其它相应性能指标。在下一步工作中,将对信用数据增广技术进行改进,研究改进后的数据增广技术与机器学习算法相融合,以进一步提高个人信用评分模型性能。

参考文献:

[1] 石庆焱,靳云汇. 多种个人信用评分模型在中国应用的比较研究[J]. 统计研究,2004(6):43-47.
[2] 刘峙廷. 我国 P2P 网络信贷风险评估研究[D]. 南宁:广西大学,2013.
[3] 秦宛顺. 一个基于 Logistic 回归的个人信用评分模型[C]. 中国数量经济学会,2003.
[4] 李建平,徐伟宣,石勇. 基于主成分线性加权综合评价的信用评分方法及应用[J]. 系统工程,2004(8):64-68.
[5] 金妍彦. 遗传规划模型在我国个人信用评估中的应用研究[D]. 哈尔滨:哈尔滨工业大学,2006.
[6] 徐少锋. FISHER 判别分析在个人信用评估中的应用[J]. 统计与决策,2006(2):133-135.
[7] 王静,王延清,何德权. 基于多层前馈神经网络的个人信用评分模型[J]. 经济师,2004(12):20-21.
[8] 肖文兵,费奇,万虎. 基于支持向量机的信用评估模型及风险评价[J]. 华中科技大学学报(自然科学版),2007(5):23-26.
[9] 萧超武,蔡文学,黄晓宇,等. 基于随机森林的个人信用评估模型研究及实证分析[J]. 管理现代化,2014,34(6):111-113.
[10] 朱兵,贺昌政,李慧媛. 基于迁移学习的客户信用评估模型研究[J]. 运筹与管理,2015,24(2):201-207.
[11] 殷爽,姜明辉. 基于 PSO 的个人信用评估组合预测模型[J]. 经济研究导刊,2008(14):83-86.
[12] 朱毅峰,孙亚南. 精炼决策树模型在个人信用评估中的应用[J]. 统计教育,2008(1):5-7.
[13] 向晖,杨胜刚. 个人信用评分关键技术研究的新进展[J]. 财经理论与实践,2011,32(4):20-24.
[14] 肖进,刘敦虎,顾新,等. 银行客户信用评估动态分类器集成选择模型[J]. 管理科学学报,2015,18(3):114-126.
[15] 陈力,黄艳莹,游德创. 一种基于 Boosting 的集成学习算法在银行个人信用评级中的应用[J]. 价值工程,2017,36(18):170-172.
[16] 卢海涛,吴磊,周建云,等. 基于 Faster R-CNN 及数据增广的满文文档印章检测[J]. 大连民族大学学报,2018,20(5):455-459.
[17] 蒋梦莹,林小竹,柯岩. 基于优化分类的数据增广方法[J]. 计算机工程与设计,2018,39(11):3559-3563.
[18] 王钰清,陆文凯,刘金林,等. 基于数据增广和 CNN 的地震随机噪声压制[J]. 地球物理学报,2019,62(1):421-433.

(责任编辑:孙　娟)