文章编号：2096-1472(2018)-06-29-03
DOI:10.19644/j.cnki.issn2096-1472.2018.06.009

基于XGBoost的信用风险分析的研究

赵天傲，郑山红，李万龙，刘 凯
(长春工业大学计算机科学与工程学院，吉林 长春 130012)

摘 要：在大数据时代如何利用数据挖掘处理海量数据从而对信用风险进行预测分析成为了当下非常重要的问题，本文运用XGBoost算法建立信用风险分析模型，运用栅格搜索等方法调优XGBoost参数，基于以AUC、准确率、ROC曲线等评价指标，与决策树、GBDT、支持向量机等模型进行对比分析，基于德国信用数据集验证了该模型的有效性及高效性。

关键词：信用风险分析；XGBoost；数据挖掘；栅格搜索
中图分类号：TP39 **文献标识码**：A

A Study of the Credit Risk Analysis Based on XGBoost

ZHAO Tianao,ZHENG Shanhong,LI Wanlong,LIU Kai
(*School of Computer Science & Engineering,Changchun University of Technology,Jilin* 130012,*China*)

Abstract: How to use data mining to process massive data to predict and analyze credit risks has become a very important issue in the era of big data.This paper adopts the XGBoost algorithm to establish a credit risk analysis model,and uses grid search and other methods to tune the XGBoost parameters based on AUC.The evaluation indicators such as accuracy rates,ROC curves,etc.are compared with the models such as decision tree,GBDT,and support vector machine.The validity and efficiency of the model are verified based on the German credit data set.

Keywords:credit risk analysis;XGBoost;data mining;grid search

1 引言(Introduction)

银行信用风险的大小和质量决定着银行盈利水平的高低，对银行稳定、长远的发展有着至关重要的影响[1]，银行使用数据挖掘方法建立目的明确、层次分明的信用风险分析模型有着重要价值。

早期的信用风险研究寻求数学解决方法，Z分数模型等都是比较具有代表性的方法[2,3]。随着银行信贷的大规模增长及客户信用信息的迅速变化，形成了复杂的数据资源，信用风险的形式与日俱增。因此，Hashemi and Blanc、Guilherme Barreto Fernandes、谢宇等分别采用神经网络和粗糙集成分集合[4]、logistic模型作为解释变量[5]、改进BP人工神经网络模型[6]对银行信用风险进行预测得到了较好改进。但以上的方法在预测精度和准确性上还有待提高。

本文引入XGBoost(eXtreme Gradient Boosting)[7]算法建立信用风险分析优化模型，基于UCI上德国信用数据集与决策树、GBDT、支持向量机等模型进行对比分析，验证了XGBoost模型应用于信用风险分析具有更好的性能。

2 XGBoost介绍(Introduction to XGBoost)

XGBoost由陈天奇博士提出的boosting型树类算法，能进行多线程并行计算，通过一次次迭代生成一代代新的树，实际上是把很多分类性能较低的弱学习器组合成一个准确率高的强学习器，每个决策树可能没有良好的分类效果，但是多个分类的结果肯定会得到更准确的预测。XGBoost加入正则项到目标函数寻求最优解，平衡目标函数的下降和模型的复杂度，避免出现过拟合现象，具有运行速度快、分类效果好、支持自定义损失函数等优点。我们希望建立K棵使树群的预测值尽量真实且泛化能力强的回归树。

XGBoost最根本就是由决策树集成而来，我们把树模型写成：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F \tag{1}$$

式(1)中，$F=\{f(x)=w_{q(x)}\}$，其中F对应所有回归树的集合，x_i表示第i个特征向量，每个f是树空间F的一棵树，每一棵树f_k对应一个独立的叶子权重w和树结构q。此时需要引入目标函数：

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta) \quad (2)$$

L部分为误差函数，表示模型拟合数据的程度，Ω表示正则项，是所有正则化项累加和，用来处理复杂模型，对复杂模型进行简单化处理。对于模型误差部分用additive training训练，通过对平方误差泰勒展开二次项，带入正则化项$\Omega(f) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$，得到最终目标函数为：

$$Obj(t) = \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)w_j^{\,2}] + \gamma T$$

其中的γ和λ是XGBoost自定义的，显然γ、λ越大，表示希望获得更简单的树，这样处理能很清楚的理解这个目标，*Obj*分数越小表示生成的树的结构越好。至此树的类型已经能够确定，接下来需要进行树的分裂，采用贪心生长树的方法，遍历所有特征，从而找到最优的特征分裂，到达一定深度或不能再分裂时停止，基于目标函数值比较分裂前后的最小目标函数值，增益最大的点为最优点，对应特征为最优特征。

3 基于XGBoost的预测方法(The prediction method based on XGBoost)

3.1 数据预处理

本文使用的数据来自UCI上公开的德国信用数据集，包括24个变量。获得该数据集后，首先标准化处理数据，清理数据集中的异常值，纠正错误数据，通过平滑噪声、数据规约等方式使得数据更加适用于本模型，同时添加ID属性，对每个属性添加属性名并做规范化。

3.2 XGBoost的参数优化

本文对于XGBoost涉及优化的参数有：max_depth、min_child_weight、gamma、seed、objective。

max_depth表示树的最大深度，能够避免过拟合，限制树分裂的程度，值越大，模型越容易产生局部最优情况，典型值3-10；min_child_weight确定最小叶节点样本权重和，值较大能够避免局部特殊样本的学习，但是值过高会导致欠拟合。由于以上三种参数值都为整数值，且值的范围较小，所以运用栅格搜索法进行调整寻求最优参数值，栅格搜索法是一种穷举搜索方法，它指定参数值，排列每个参数的可能值，列出所有可能的组合，并生成“网格”，然后训练每个组合，进行交叉验证评估性能。

gamma指所需的最小损失函数下降值，满足该值节点才会分裂，值越大，算法越保守；seed是随机种子个数，用于调整参数、显示随机数据结果。因为两组参数的值为随机值或者连续值，所以随机选取几个合理的数值分别进行调整，选取最优的数值作为参数值。

objective定义需要最小化的损失函数，常用值有：二分类逻辑回归—binary：logistic；多分类器—multi：softmax。本文为二分类数据，根据经验值来确定参数。

3.3 XGBoost模型描述

XGBoost最根本的就是希望建立K棵回归树，使得准确率高、泛化性好、预测误差尽量小，叶子节点尽量少的目标函数才能训练出更好的模型，利用贪心策略及二次最优化确定最优节点及最小的损失函数，以此为依据进行树分裂，得到小树苗，接下来按照上述方式继续分裂，并继续形成新树，根据之前的预测每次都会建立最优的树，当达到max_depth时停止迭代；此时我们得到了最基本的模型，之后使用栅格搜索等方法对几种参数进行优化，从而分析数据。

优化后的XGBoost模型如下，模型图如图1所示：

(1)初始化回归树$f(x)$，损失函数集合$l_t(x)$，此时模型为常数值

(2)While k on 1,2,3,...,T do

(3)do

(4)计算损失函数L的最小值m

(5)把m加到$l_t(x)$中

(6)While t on 1,2,3,...,T

(7)$l_t(x)$中选取m最小时t对应的$f_t(x)$开始建树

(8)采用贪心法寻找最优分裂节点迭代生成新的树$f_t'(x)$

(9)If deep>max deep break

(10)得到最终模型

(11)利用Raster search等方法调优参数

(12)优化模型分析数据

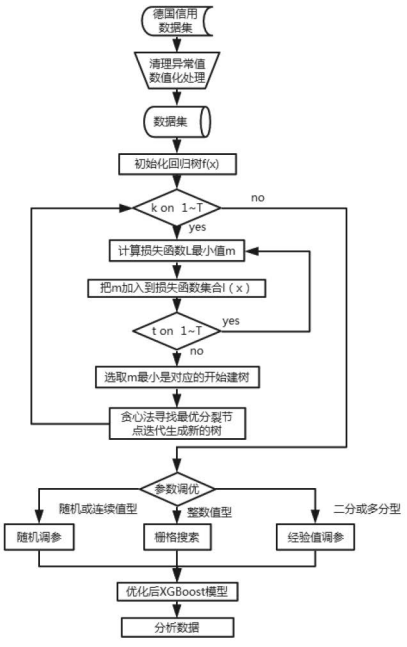

图1 XGBoost模型

Fig.1 XGBoost model

其中，贪心法寻找最优分裂节点算法如下：

输入：样本集I，维度d

(1)$G \leftarrow \sum_{i \in I_j} g_i, H \leftarrow \sum_{i \in I_j} h_i$

(2)for k=1 to d do

(3)$G_L \leftarrow 0, H_L \leftarrow 0$

(4)for j in sorted(I,by x_{kj}) do

(5)$G_L \leftarrow G_L + g_j, H_L \leftarrow H_L + h_j$

(6)$G_R \leftarrow G - G_L, H_R \leftarrow H - H_L$

(7)$score \leftarrow \max(score, \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda})$

(8)end for

(9)end for

4 实验分析(Experimental analysis)

本文所使用的德国信用数据集如表1所示。

表1 德国信用数据集

Tab.1 German credit data set

属性名称	属性描述	类型(举例)
Status of existing checking account	现有支票帐户的状态	A11、A12、A13、A14
Duration in month	月期限	numerical
Credit history	信用记录	A30、A31、A32、A33、A34
Purpose	目的	A40、A41、A42、A43…
Credit amount	信用额	numerical
Savings account/bonds	储蓄账户/债券	A61、A62、A63、A64、A65
Present employment since	目前就业情况	A71、A72、A73、A74、A75
…	…	…
foreign worker	外国工人	Yes、No

使用python语言和Pycharm软件来实现模型，用到了pandas、itertools、numpy等包，分别使用决策树、GBDT、SVM及XGBoost进行分析比较，采用K折交叉验证的方式(5折、10折交叉验证)分别处理数据集，对比几种算法的精密度Precision(Precision=TP/(TP+FP))、召回率Recall(Recall=TP/(TP+FN))、准确度AUC值、F1值(F1 Score=P*R/2(P+R)、Accuracy(Accuracy=(TP+TN)/(TP+FP+TN+FN))、真假阳性率False Positive Rate—True Positive Rate折线图；P和R分别为Precisionl和Recal)等指标，其中TP为真阳性，FP为假阳性，TN为真阴性，FN为假阴性。

表2 德国信用数据集上准确性和AUC值的评估效果(%)

Tab.2 The evaluation effect of accuracy and AUC value on German credit data set(%)

	评估类别	DT	GBDT	SVM	XGBoost	优化后XGBoost
5折	Accuracy	82.67	86	85.07	87.33	89.6
	AUC Score	89.56	91.84	91.50	92.98	95.45
10折	Accuracy	83.07	87	85.6	89.33	90.67
	AUC Score	90.96	91.9	92.16	95.20	95.97
平均	Accuracy	82.87	86.5	85.34	88.33	90.14
	AUC Score	90.26	91.87	91.83	94.09	95.17

表3 在德国信用数据集上精确度、回收率和F1值的评估效果(%)

Tab.3 Evaluation of accuracy,recovery rate and F1 value on German credit data sets(%)

	评估类别	DT	GBDT	SVM	XGBoost	优化后XGBoost
5折	Precision	82	86	85	87	90
	Recall	83	86	85	87	90
	F1-score	82	85	84	87	90
10折	Precision	83	87	86	89	91
	Recall	83	87	86	89	91
	F1-score	82	87	85	89	90
平均	Precision	82.5	86.5	85.5	88	90.5
	Recall	83	86.5	85.5	88	90.5
	F1-score	82	86	84.5	88	90

从表2和表3可以看出，XGBoost比决策树、GBDT、SVM在各项指标上的值均有不同程度的提高；同时，优化后的XGBoost在各项指标上都有所提升；优化后的XGBoost的平均Accuracy和AUC值比决策树、GBDT、支持向量机分别高出4.19%、3.3%、3.34%；优化后的XGBoost平均F1值比决策树、GBDT、支持向量机分别高出7.5%、3.5%、5%；相比较其他几种算法，准确性、召回率均有提高。

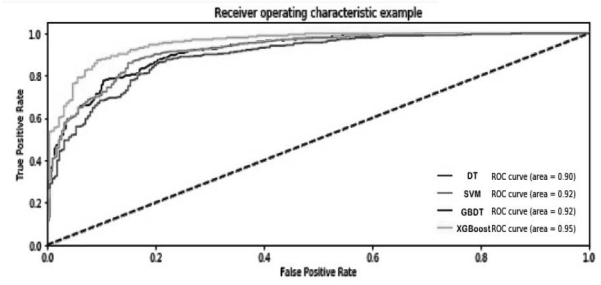

图2 ROC曲线比较图

Fig.2 ROC curve comparison graph

图2为XGBoost、决策树、GBDT、SVM的ROC曲线图

(受试者工作特征曲线)。曲线下方的面积即为AUC，当AUC越接近1时，分类器越完美；由图可知，XGBoost的ROC曲线最优，这说明XGBoost的分类效果最好。

5 结论(Couclusion)

本文研究基于XGBoost算法对信用风险进行分析，以德国信用公开数据集作为数据源，采用K折交叉验证法，通过栅格搜索、经验值调参等方法对参数进行调整，基于AUC、准确率、ROC曲线等评价指标，与决策树、GBDT、支持向量机等模型进行对比分析。实验表明调优后的XGBoost算法应用于数据集上比调参前在各方面均有明显调整，同时XGBoost算法相对于常用的决策树、GBDT和支持向量机算法，无论是准确性还是分类效果等方向都有更加明显的优势，验证了XGBoost模型的有效性和精确度。

参考文献(References)

[1] Cheng-Lung Huang,Mu-Chen Chen,Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines[J].Expert Systems with Applications,2007,33(4):847-856.

[2] Edward I Altman,Anthony Saunders.Credit risk measurement: Developments over the last 20 years[J].Journal of Banking and Finance,1997,21(11):1721-1742.

[3] Michel Crouhy,Dan Galai,Robert Mark.A comparative analysis of current credit risk models[J].Journal of Banking and Finance,2000,24(1):59-117.

[4] R.R.Hashemi,L.A.Le Blanc,C.T.Rucks,A.Rajaratnam. A hybrid intelligent system for predicting bank holding structures[J].European Journal of Operational Research,1998,109(2):390-402.

[5] Guilherme Barreto Fernandes,Rinaldo Artes.Spatial dependence in credit risk and its improvement in credit scoring[J].European Journal of Operational Research,2016,249(2):517-524.

[6] 谢宇.基于人工神经网络的商业银行信贷风险预警研究[D].暨南大学,2010.

[7] Chen T,Guestrin C.XGBoost:A Scalable Tree Boosting System[J].KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2016:785-794.

作者简介：

赵天傲(1993-)，男，硕士生.研究领域：数据挖掘.

郑山红(1970-)，女，博士，教授.研究领域：软件工程.

李万龙(1963-)，男，教授.研究领域：软件工程.

刘　凯(1991-)，女，硕士生.研究领域：人工智能.

(上接第35页)

析，如图7所示。

图7 数据查询

Fig.7 Data query

5 结论(Conclusion)

课堂教学是素质教育的主要阵地，是教师和学生教与学活动相互的直接表现。课堂行为动作监控可以全方位跟踪学生学习，本文智能课堂监控与分析系统充分结合了智能识别技术、计算机技术和网络技术，为教学质量提供真实有效的依据。本文所涉及的人工智能机器识别中的行为识别作为当下最热门的技术之一，应用层面广博。本文的研究和设计成果可以为各个高校课堂学生行为管理提供有益的借鉴和参考。但是行为识别程序在实际应用时，识别准确率还有待提高，需要进一步研究完善。

参考文献(References)

[1] 樊凌,戴雯惠,唐寅.基于智能跟踪的课堂教学监控系统的探索与研究[C].International Conference on Education Technology & Training,2010.

[2] 黄健荣.梧州学院课堂教学质量监控系统的设计与实现[D].电子科技大学,2013.

[3] 徐金凤,张路遥.高职院校教学质量监控信息化平台的设计——以无锡职业技术学院为例[J].无锡职业技术学院学报,2017,16(3):41-44.

[4] 李霞.校园视频监控系统的研究与实现[D].山东大学,2015.

[5] 袁国武.智能视频监控中的运动目标检测和跟踪算法研究[D].云南大学,2012.

[6] 荆洲,权伟,唐杰,等.基于人脸识别的智能课堂点名系统[J].软件工程,2017,20(5):43-46.

[7] 王云良,汤慧芹,顾卫杰.高职院校智能化课堂教学质量监控系统研究[J].软件工程,2015(3):28-29.

作者简介：

戴振泽(1998-)，男，本科生.研究领域：物联网工程.

施　艳(1998-)，女，本科生.研究领域：物联网工程.

郑少伟(1998-)，男，本科生.研究领域：软件工程.

郭梓文(1999-)，男，本科生.研究领域：图像处理.

丁王斌(1990-)，男，硕士，助教.研究领域：软件工程.本文通讯作者.