[文章编号] 1007－7405(2021)01－0089－08 DOI:10.19715/j.jmuzr.2021.01.13

基于不平衡数据的个人信贷违约测度探索

郭 畅

（安徽大学经济学院，安徽 合肥 230601）

[摘要] 针对个人信贷风险评估中存在的类别不平衡问题，为了提升信贷违约客户的识别能力，提出基于欠采样改进的集成模型。该模型从“数据”层面进行批量欠采样处理，从“算法”层面对现有的集成模型进行再次集成。在UCI台湾信用卡信贷数据集上，结合模型整体效果的测度AUC值、精度方面的测度F_1值和区分度指标KS值进行评估。结果表明，基于欠采样改进的Batch-US-集成模型总能取得更优结果，其中属Batch-US-Xgboost模型最优，接着对Batch-US-集成模型的子模型个数和模型有效性进行探索，证实了子模型个数并非越多越好的结论。改进后的Batch-US-集成模型能够有效提升信贷风险违约预测的效果。

[关键词] 信贷风险；违约预测；类别不平衡；集成模型；子模型个数

[中图分类号] TP 391；C 81

Personal Credit Default Measurement Research Based on the Imbalanced Data

GUO Chang

(School of Economics, Anhui University, Hefei 230601, China)

Abstract: Aiming at the problem of category imbalance in personal credit risk assessment, in order to improve the identification ability of credit default customers, an improved integration model based on under sampling is proposed. The model is based on under sampling processing from the data level and reintegrating the existing integration model from the algorithm level, and studies the improvement effect of this model. On the UCI Taiwan credit card credit data set, it evaluates the AUC value of the overall effect of the model, the F_1 value of the accuracy and the KS value of the differentiation index . The results show that the Batch-US-Ensemble models based on the under sampling process can always achieve better results, and the Batch-US-Xgboost model is the best among all ensemble models. Then, the number of sub models and the validity of Batch-US-ensemble models are explored, which proves that the number of sub models is not the more the better. The improved Batch-US-Ensemble model can effectively improve the effect of credit risk default prediction.

Keywords: credit risk; default prediction; class imbalance; ensemble model; the number of sub models

0 引言

信贷风险一直是商业银行不可避免的信用风险之一，然而信用风险管控对风险评级模型有较高的要求，2019年来，随着数字普惠金融模式“开放银行+”的推进，带来了个人、小微信贷业务的提升[1]。此时，随着数据量的快速增多，如何对抗不平衡数据的弊端，建立更加精确的信用风险违约

[收稿日期] 2020－03－31

[基金项目] 国家自然科学基金面上项目（NSFC11871074）

[作者简介] 郭畅（1997—），女，硕士生，从事数据挖掘方向研究。E-mail:trickleguo1021@163.com

预测模型，降低商业银行所遭受的客户信贷风险，在当今金融科技浪潮下更凸显其重要意义。信贷违约预测的目标是提前预知哪些客户更倾向于违约。然而违约事件的发生是少数的，收集到的信贷数据往往呈现出正负样本分布不平衡的结构。常用的机器学习算法往往建立在训练集各个类别数目分布比例大致均等的假设上[2-3]，而在非平衡数据集中的表现一般较差。因此，如何处理不平衡的信贷数据集对风控模型精度的提升显得尤为重要。

随着人工智能第三次浪潮来袭，以神经网络、支持向量机和集成方法为首的机器学习算法越来越多地涌入信用风险评估领域。陈力[4]通过综合不同的采样方法，并结合集成学习算法和模型评价指标构建新的算法模型 RHSBoost，将“数据”和“算法”两个方面结合来对银行信用评级的不同数据集进行分类预测并得到了不错的效果。古平等[5]在结合“数据”和“算法”的层面上提出 AdaBoost-SVM-MSA 算法，按照一定规则将 SVM 分错的样本划分为噪声样本、危险样本、安全样本三种类型，然后直接删除噪声样本，取安全样本进行 SMOTE 过采样，显著提高了模型分类准确率。董路安等[6]在文献［5］的基础上，运用“安全样本”消除噪声干扰，并将 Weight-SMOTE 方法应用于决策树模型中，提升了信用评级模型的可解释性，但对正负样本均进行同原数据比例的 SMOTE 抽样却忽略了信用评估数据的不平衡结构。李毅等[9]分别采取过采样[7]、欠采样[7]、SMOTE 人工合成[8]的三种方法得到三个数据集，对处理后的三个数据集分别建立三个机器学习模型，并与未处理数据的三种模型结果进行对比试验，得出过采样结合随机森林模型评估的结果高于其他模型。陈启伟等[10]从欠抽样方法入手，从多数类样本中反复抽取和少数类样本量已知的样本组成多个子数据集，对多个数据子集建立模型并采用简单平均集成得到较好的预测性能。然而，文献［7－9］未讨论现有欠抽样方法上的改进效果，文献［10］未从子模型个数和模型评价效果方面进行研究。

结合上述文献的不足，本文同时从“数据”的修正和“算法”的改进入手，选择 UCI 真实业务场景的 30 000 条记录 23 个指标的台湾客户信用卡信贷数据，将“数据”和“算法”两个层面改进的 Batch-US-RF 集成模型、Batch-US-Xgboost 集成模型与 Batch-US 处理后的单模型、未经 Batch-US 处理的单个集成模型，与单模型进行对比，并研究模型在不平衡信用卡信贷数据上的违约预测效果。

1 方法与模型

1.1 Batch-US-集成模型

批量欠采样（Batch-US）是基于随机欠采样（random under sample）方法造成的多数类样本信息缺失的改进，它对多数类样本采取多次欠采样，再和少数类样本组合成一系列新样本，来消除由于信息缺失带来的分类器分类效果不稳定的缺陷。首先，使用欠采样将多数类样本划分为多个部分，其中每部分与少数类样本数相同；接着，将这些数据和所有少数类样本组成新的子集；然后，对不同的训练子集建立差异化的集成模型；最后，将每折交叉验证的预测集预测其概率并进行简单算数平均后再组合。算法的整体结构见图 1 所示，其中本文训练的子模型分别选择随机森林和 Xgboost，将所有子模型的输出概率的平均作为分类结果输出。

输入：数据集 $D = \{(x_i, y_i), i = 1,2,\cdots,N, y_i \in \{0,1\}\}$。0 类（多数类）样本数记为 N_m，1 类（稀有类）样本数记为 N_s，有 $N_m + N_s = N$。

输出：$H(x) = (1/k)\sum_{j=1}^{k} h_j(x)$，其中 k 表示子模型的个数，$h_j(x)$ 表示第 j 个基分类器。

算法步骤：

1）将数据集 D 中的 0 类样本和 1 类样本分别记为 S_m 和 S_s，$k = \text{ceil}(S_m/S_s)$ 进一取整；

2）for $j = 1, 2, \cdots, k$, do；

3）　从 1 ~（$N_s - i + 1$）中随机抽样，取出对应序号的样本 x'；

4）　在类 0 样本中取出所选样本 $S_s = S_s - x'$；

5）　随机欠采样后的数据集 $\{D_j' = (x_i, y_i), i = 1,2,\cdots,N - S_s \cdot R_S/(R_S + 1), j = 1,2,\cdots,k, y_i \in \{0,1\}\}$，$R_S$ 表示采样比率；

6）　对每个 D_j' 训练一个子模型，记 $h_j(x)$；

7）end for；

8）对每个分类器测试集预测结果进行集成得到最终概率：$H(x) = (1/k)\sum_{j=1}^{k} h_j(x)$。

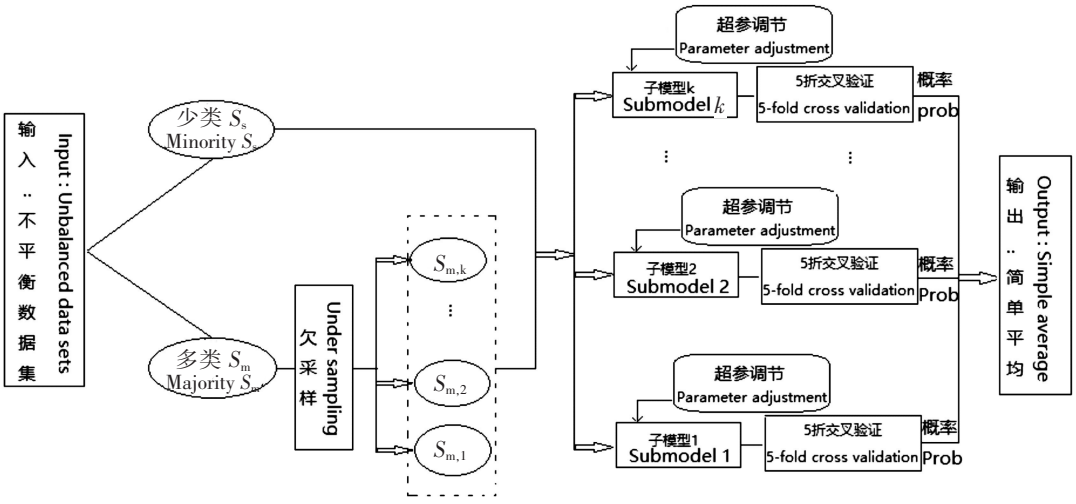


图 1　Batch-US-集成模型整体结构图

Fig.1　Overall structure of Batch-US-Ensemble model

1.2　子模型确定

1.2.1　随机森林模型

集成学习模型有两个重要的方面——基于 Bagging 的集成模型和基于 Boosting 的集成模型。基于 Bagging 的集成模型是将多个有差异的分类器取平均，能够解决一定程度上的模型不稳定问题。随机森林（random forest，RF）作为典型的 Bagging 类模型，可和采样技术结合被用于解决类不平衡问题。本文就是利用样本采样技术构造平衡随机森林[11]，并对随机森林的预测结果再次组合。

随机森林是基于 Bagging 的集成学习方法，它采用 bootstrap 自助抽样从数据集中抽取多个子样本，对抽样后的子样本分别建立具有差异性的 CART 决策树模型（每个模型随机选取 m 个特征，本文选择使模型误差最小的 m），最后对每个分类器的预测结果进行组合，组合方法采用多数表决（投票法），算法的流程如图 2 所示。

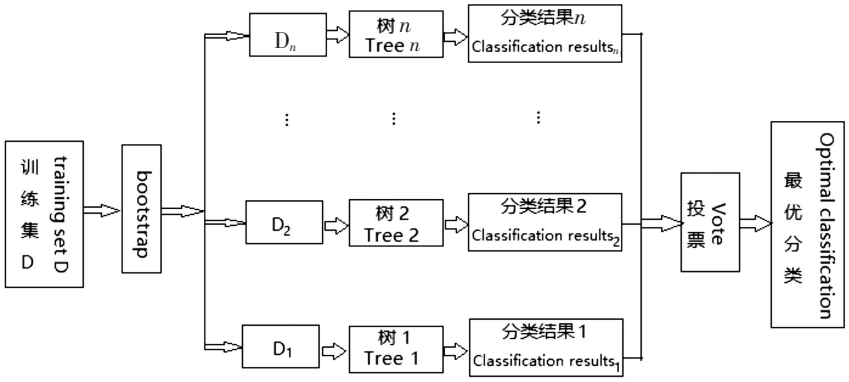


图 2　随机森林算法流程图

Fig.2　Flow chart of random forest algorithm

1.2.2 极限梯度提升模型

基于 Boosting 的集成模型 Xgboost[12]使用贪心算法和加法模型，每次构建一个当下最优的树模型，将所有树模型的最终结果求和作为最终的预测结果。其优点在于 GBDT 算法的求解采用了二阶梯度，并加入了正则化项，由于树模型容易过拟和，因此通过同时控制模型损失函数和模型复杂度得到更优结果。模型的原理和推导见文献［10］。当基模型同样选择树模型时算法的流程如图 3 所示。

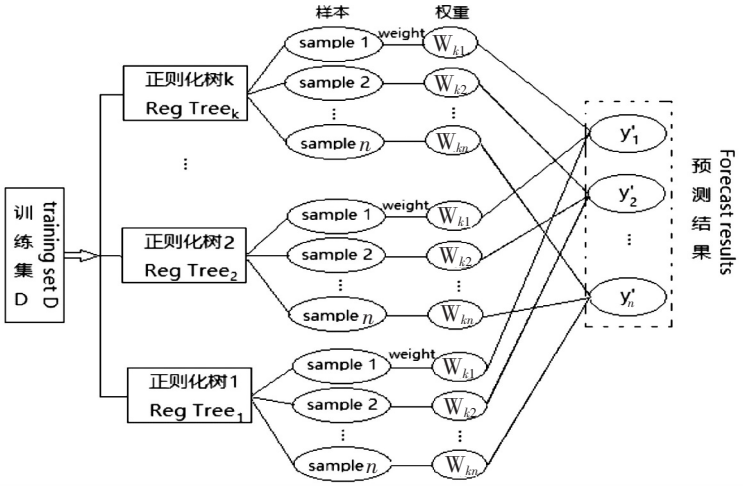


图 3 Xgboost 算法流程图

Fig.3 Xgboost algorithm flow chart

2 研究设计

2.1 指标类型

本文数据源于 UCI 机器学习网站（http://archive.ics.uci.edu/）公布的台湾客户信用卡信贷数据集，3 万条样本数据包括来自三个方面用户信息的 23 个指标数据。其中：正常客户占数据的 77.88%，违约客户占 22.12%；人口统计学特征的用户基本属性信息指标包括性别、年龄、教育程度、婚姻状况等 4 个变量；金融特征的借贷相关信息指标包括月还款情况、月账单、月支付金额等 19 个字段。由于数据存在错误值和离群值。对数据进行简单预处理后，具体的数据说明见表 1。

表 1 变量说明表

Tab.1 Variable description

指标类型 Index type		变量名称 Variable name	变量解释 Variable explanation	取值范围 Value range	变量类型 Variable type
因变量 Dependent variable		target	违约与否 Breach of contract or not	0－正常,1－违约 0－good,1－bad	分类型 Categorical
自变量 Independent variable	用户基本属性信息 User basic attribute information	sex	性别 Gender	1－男,2－女 1－male,2－female	分类型 Categorical
		age	年龄 Age	［21,79］	数值型 Numerical
		education	教育程度 Education	1－研究生,2－本科,3－高中,4－其他 1－graduate,2－undergraduate,3－high school,4－others	分类型 Categorical
		marriage	婚姻状况 Marital status	1－已婚;2－单身;3－其他 1－married;2－single;3－others	分类型 Categorical

续表

指标类型 Index type		变量名称 Variable name	变量解释 Variable explanation	取值范围 Value range	变量类型 Variable type
自变量 Independent variable	征信相关变量 Credit related variables	limit_bal	信用额度 Credit line	[10 000,1 000 000]	数值型 Numerical
		pay_0	第 6 月还款情况 Repayment in the sixth month	-2 ~ 8	分类型 Categorical
		pay_2	第 5 月还款情况 Repayment in the fifth month	-2 ~ 8	分类型 Categorical
		pay_3	第 4 月还款情况 Repayment in the fourth month	-2 ~ 8	分类型 Categorical
		pay_4	第 3 月还款情况 Repayment in the third month	-2 ~ 8	分类型 Categorical
		pay_5	第 2 月还款情况 Repayment in the second month	-2 ~ 8	分类型 Categorical
		pay_6	第 1 月还款情况 Repayment in the first month	-2 ~ 8	分类型 Categorical
		bill_amt1	第 6 月账单 6th month bill	[-165 580,964 511]	数值型 Numerical
		bill_amt2	第 5 月账单 5th month bill	[-69 777,983 931]	数值型 Numerical
		bill_amt3	第 4 月账单 4th month bill	[-157 264,1 664 089]	数值型 Numerical
		bill_amt4	第 3 月账单 3rd month bill	[-170 000,891 586]	数值型 Numerical
		bill_amt5	第 2 月账单 2nd month bill	[-81 334,927 171]	数值型 Numerical
		bill_amt6	第 1 月账单 1st month bill	[-339 603,961 664]	数值型 Numerical
		pay_amt1	第 6 月支付金额 Payment amount in the 6th month	[0,873 552]	数值型 Numerical
		pay_amt2	第 5 月支付金额 Payment amount in the 5th month	[0,1 684 259]	数值型 Numerical
		pay_amt3	第 4 月支付金额 Payment amount in the 4th month	[0,896 040]	数值型 Numerical
		pay_amt4	第 3 月支付金额 Payment amount in the 3rd month	[0,621 000]	数值型 Numerical
		pay_amt5	第 2 月支付金额 Payment amount in the 2nd month	[0,426 529]	数值型 Numerical
		pay_amt6	第 1 月支付金额 Payment amount in the 1st month	[0,528 666]	数值型 Numerical

注：pay_0 ~ pay_6 取值为 -2 表示没有支出，取值为 -1 表示全额偿还，等于 0 表示客户按时分期还款，取值大于 0 表示存在不同程度的拖欠。由于逾期超过 3 月的客户数较少，因此后续将其合并为逾期 3 月以上。

Notes: pay_ 0 ~ pay_ 6, a value of -2 means no expenditure, a value of -1 means full repayment, equal to 0 means timely repayment by installments, a value greater than 0 means there are different degrees of arrears. Due to the small number of customers overdue for more than 3 months, they are subsequently merged into overdue for more than 3 months.

2.2　模型建立与评价

2.2.1　评价指标

对于本文正负样本比例约 3.5∶1 的不均衡的数据集，传统的基于准确率的模型评价指标已经不再适用[13-14]。基于此，本文选取 F_1 指标和 ROC 曲线下面积 AUC 来评价模型的预测精度，用 KS 值(kolmogorov smirnov)[15]检测实际风控模型的好坏。KS 取值越接近 1 则模型区分度越高，预测能力越强。模型评价指标由表 2 混淆矩阵计算得出，指标计算公式为：查准率 $P = N_{TP}/(W_{TP} + N_{FP})$；查全率 $R = N_{TP}/(N_{TP} + N_{FN})$；$F_1 = 2 \times \text{precision} \times \text{recall}/(\text{precison} + \text{recall})$。

表 2　二分类结果混淆矩阵

Tab. 2　Confusion matrix of classification results

真实情况 Actual	预测结果 Predict	
	0 类 Class 0	1 类 Class 1
0 类 Class 0	TN(正负例)	FP(假正例)
1 类 Class 1	FN(假负例)	TP(真正例)

2.2.2 模型建立和评估

由表 1 变量说明可见，本文选用的客户信用卡信贷数据间量纲差别较大，需要对数据进行标准化处理。本文为了更好地进行模型评估，增强模型稳定性，对每个模型分别进行5折交叉验证（模型如表 3 所示）。对于本文不平衡的信用卡信贷数据，经阈值调优，对未经平衡处理的数据阈值设定为 0.45，处理后的数据阈值设定为 0.55。

由于树模型容易过拟和，本文对选择的每个树模型进行参数调优（见表 4），并在 Batch-US 模型集成过程的 Rstudio 中构建 ovun.sample 随机欠采样函数，通过设置 seed 随机种子的不同，对每一折交叉验证数据构建多个随机欠采样子样本，得到多个平衡子样本，再加上参数调节，使每个子模型更具差异性，从而增加集成模型的泛化能力。其中对 Batch-US 改进的模型分别构建 10 个差异性的子模型。

表 3 模型类型及名称

Tab. 3 Model type and name

模型类型 Model type	模型名称 Model name
单模型 Single model	DT
	LR
	KNN
单个集成模型 Single ensemble model	RF
	Xgboost
单模型 Batch-US-集成 Single model Batch-US-Ensemble	Batch-US-DT
	Batch-US-KNN
	Batch-US-LR
集成模型 Batch-US-集成 Ensemble model Batch-US-Ensemble	Batch-US-RF
	Batch-US-Xgboost

表 4 树模型调优参数及范围

Tab. 4 Tuning parameters and scope of tree model

模型 Model	参数 Parameter	符号表示 Symbolic representation	取值范围 Value range
决策树 Decision tree	树最大深度 Max depth of tree	max_depth	4 ~ 5
	树复杂度 Complexity of tree	cp	0.1 ~ 0.3
随机森林 Random forest	树的个数 Number of tree	ntree	360 ~ 380
	最大特征数 Max feature number	mtry	3 ~ 5
Xgboost	学习率 Learning rate	eta	0.03 ~ 0.05
	树最大深度 Max depth of tree	max_depth	3 ~ 5
	特征采样占比 Feature sampling ratio	colsample_bytree	0.8 ~ 1
	L1 正则化参数 L1 regularization parameter	lambda	1 ~ 1.5
	L2 正则化参数 L2 regularization parameter	alpha	4.5 ~ 5
	迭代次数 Iterations	num_round	100 ~ 150

本文对文献［10］中的评价指标进行改进，基于准确率对不平衡数据的缺陷，选择用 F_1 值衡量模型精度，用 AUC 值评估模型的优劣，用 KS 值衡量模型的稳健性和风控能力。将 10 个模型经五折交叉验证后的预测指标平均，汇总至表 5。

表 5 模型结果汇总

Tab. 5 Summary of model results

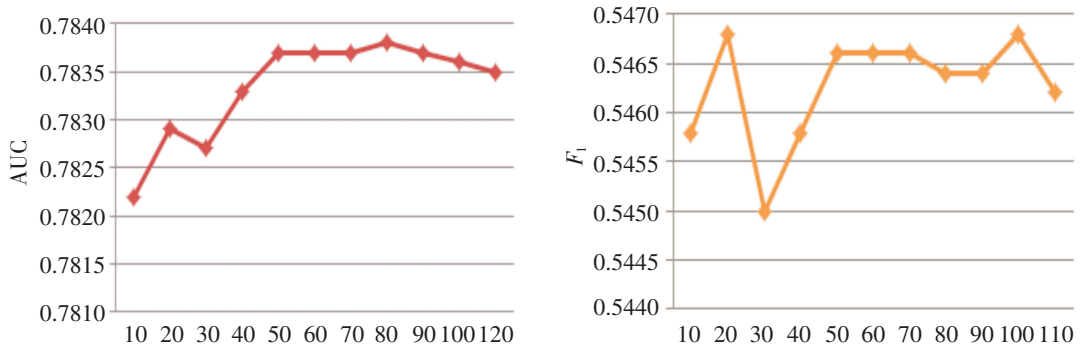
模型 Model	F_1 值 F_1-measure	AUC 值 AUC value	KS 值 KS value
DT	0.4362	0.6436	0.2873
KNN	0.4556	0.7243	0.3435
LR	0.4813	0.7670	0.4131
RF	0.5012	0.7667	0.4085
XGBoost	0.4747	0.7782	0.4288
Batch-US-DT	0.5097	0.7245	0.3903
Batch-US-KNN	0.4995	0.7443	0.3716
Batch-US-LR	0.5333	0.7679	0.4160
Batch-US-RF	0.5369	0.7796	0.4246
Batch-US-Xgboost	0.5458	0.7822	0.4354

由表 5 模型结果可知，不管是单模型还是集成模型，在通过本文的 Batch-US 批量欠采样集成后，在 F_1 值、AUC 值和 KS 值 3 个评价指标上都有明显的提升。在本身就较优的集成模型上更能进一步

提升模型的表现能力。Batch-US-RF 模型的 F_1 值、AUC 值和 KS 值分别比改进前提高了 3.57%、1.29%、1.61%；Batch-US-Xgboost 模型的 F_1 值、AUC 值和 KS 值分别比改进前提高了 7.11%、0.4%、0.66%。Batch-US-集成模型的精度衡量指标 F_1 值和 AUC 值都是 10 个模型中最优的，并且观察其区分度指标 KS 值也大于 0.4 且排名在 10 个模型中前三，说明模型风控能力较好。

表 5 评价指标结果均为本模型数量选择 $k=10$ 的结果。为了进一步研究子模型数量是否对模型精度造成影响，本文将两个 Batch-US-集成模型通过设定子模型数量 k 为 10,20,…,110 时的模型评价效果绘制学习曲线，如图 4、图 5 所示。

由图 4、图 5 可知，Batch-US-Xgboost 模型通过增加子模型数量，其 F_1 值和 AUC 值在一开始的确有一个上升幅度，但是随着模型不断增多，这三个评价指标均先趋于稳定而后随子模型个数上升甚至出现轻微下降趋势。Batch-US-RF 模型通过增加子模型数量，其 AUC 值在一开始的确有一个上升幅度，但是随着子模型不断增多 AUC 值趋于稳定；其 F_1 值在前 60 个模型的整体趋势不断上升，但是在 60 个子模型后围绕一个固定值波动（认为其趋于稳定）。因此，子模型数量并非越多越好，两个 Batch-US-集成模型的子模型数量在 60 个左右能够取得 AUC 和 F_1 指标的较优和模型较稳定的结果。

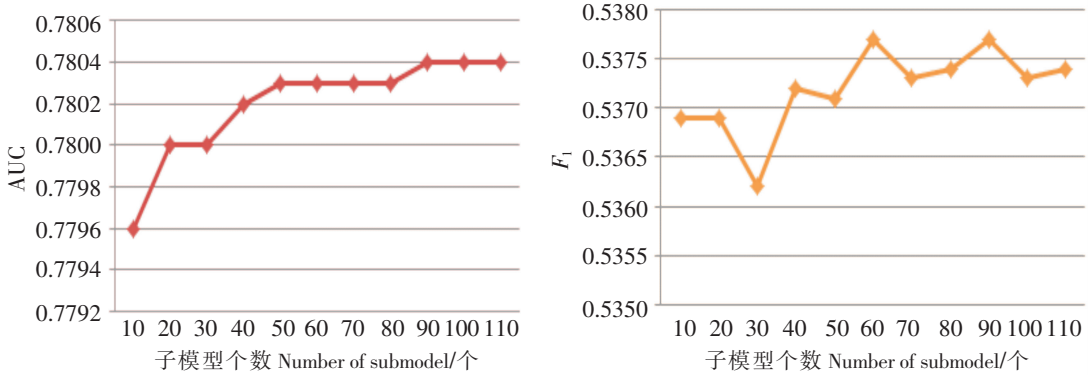


a) AUC 值的变化曲线　　b) F_1 值的变化曲线

Change curve of AUC value　　Change curve of F_1 value

图 4　Batch-US-Xgboost 模型精度指标随子模型数量变化的学习曲线

Fig.4　Learning curve of Batch-US-Xgboost model accuracy index changing with the number of submodels



a) AUC 值的变化曲线　　b) F_1 值的变化曲线

Change curve of AUC value　　Change curve of F_1 value

图 5　Batch-US-RF 模型精度指标随子模型数量变化的学习曲线

Fig.5　Learning curve of Batch-US-RF model accuracy index changing with the number of submodels

3 结论

本文使用 UCI 台湾客户信用卡信贷数据，分别对数据进行单模型、集成模型和 Batch-US 处理后建模。由表 5 可以看出，基于欠采样改进的 Batch-US-集成模型的建模结果明显优于处理之前的数据建模结果。由于在风控模型中千分之一的精度改变带来的影响也是巨大的，对不平衡数据的处理具有较大意义，本文进行 Batch-US 集成后模型的确提升了模型预测效果，且 Batch-US-集成模型总能表现出更好结果。

该系列模型从“数据”层面使用批量欠采样处理修正了随机欠采样的弊端，从“算法”层面对多个模型采用简单平均集成增加了分类器的稳定性。通过实证分析，结合模型评价指标，验证了 Batch-US-RF 和 Batch-US-Xgboost 模型不管从模型精度、综合效果方面还是从实际风控效果方面都具有较高的表现能力，尤以 Batch-US-Xgboost 模型有效性和精度最高。本文通过绘制不同子模型个数和模型评价指标的学习曲线，得出结论：对于 Batch-US-集成模型并非子模型数量越多越好，子模型的数量可以根据模型复杂度和不同评价指标的倾向性进行选择。

[参 考 文 献]

[1] 中国银行协会，普华永道会计师事务所. 中国银行家调查报告 2018 [M]. 北京：中国金融出版社，2019.

[2] 康琦，吴启迪. 机器学习中的不平衡分类方法 [M]. 上海：同济大学出版社，2017.

[3] 于化龙. 类别不平衡学习理论与算法 [M]. 北京：清华大学出版社，2017.

[4] 陈力. 银行信用评级中的不平衡分类问题研究 [D]. 广州：广东工业大学，2017.

[5] 古平，欧阳源游. 基于混合采样的非平衡数据集分类研究 [J]. 计算机应用研究，2015，32(2)：379-418.

[6] 董路安，叶鑫. 基于改进教学式方法的可解释信用风险评价模型构建 [J]. 中国管理科学，2020，28(9)：45-53 .

[7] CHAWLA N，BOWYER K W，HALL LO. SMOTE：synthetic minority over sampling technique [J]. Journal of Artificial Intelligence Research，2002，16：321-357.

[8] HAN H，WANG W Y，MAO B H. Borderline-SMOTE：a new over-sampling method in imbalanced data sets learning [C] //Proceedings of the 2005 International Conference of Intelligent Computing. Hefei：Lecture Notes in Computer Science，2005：878-887.

[9] 李毅，姜天英，刘亚茹. 基于不平衡样本的互联网个人信用评估研究 [J]. 统计与信息论坛，2017，32(2)：84-90.

[10] 陈启伟，王伟，马迪，等. 基于 Ext-GBDT 集成的类别不平衡信用评分模型 [J]. 计算机应用研究，2018，35(2)：421-427.

[11] 田臣，周丽娟. 基于带多数类权重的少数类过采样技术和随机森林的信用评估方法 [J]. 计算机应用，2019，39(6)：1707-1712.

[12] 吴金旺，顾洲一. 基于非平衡样本的商业银行客户信用风险评估——以 A 银行为例 [J]. 金融理论与实践，2018(7)：51-57.

[13] 夏利宇，何琬. 信用评级模型构建的统计学解读 [J]. 征信，2019，37(6)：44-48.

[14] 刘志惠，黄志刚，谢合亮. 大数据风控有效吗？——基于统计评分卡与机器学习模型的对比分析 [J]. 统计与信息论坛，2019，34(9)：18-26.

[15] CHEN T Q，GUESTRIN C. XGBoost：a scalable tree boosting system [C] //Proc of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM，2016：785-794. DOI：10.1145/2939672.2939785.

（责任编辑 朱雪莲 英文审校 黄振坤）