第 38 卷第 3 期　　重庆工商大学学报(自然科学版)　　2021 年 6 月
Vol. 38　No. 3　　J Chongqing Technol & Business Univ(Nat Sci Ed)　　Jun. 2021
doi:10. 16055/j. issn. 1672-058X. 2021. 0003. 012

非均衡数据的债券违约预警研究

程建华, 徐恒宇*
(安徽大学 经济学院,合肥 230601)

摘　要:将上海交易所和深证交易所发行的 30 只违约债券和 468 只未违约债券作为研究样本,将债券是否违约设定为一个二分类问题进行识别分析,针对该问题构建了基于 SVM 的 ADmR-AdaboostSVM 分类模型;从企业资本结构、盈利能力、现金流量、偿债能力 4 个评估因素中筛选 16 个预警指标,运用 ADASYN 方法进行过采样合成新样本点,将特征提取 *m*RMR 方法引入债券违约领域,得出长期负债率、资本收益率、成本费用利润率以及股权比例这 4 个变量作为债券违约的最终预警指标,在此基础上运用 AdaboostSVM 模型进行风险识别。研究结果表明:在建模过程中克服了样本非均衡化问题使得分类精度显著提高,同时通过解决高维数据冗余问题,识别违约债券的准确率进一步提高,反复验证表明该模型具有较强的稳健性和有效性,具有一定的应用价值。

关键词:债券违约;ADASYN 算法;*m*RMR 算法;AdaboostSVM
中图分类号:F832. 5　　**文献标志码**:A　　**文章编号**:1672-058X(2021)03-0086-08

0　引　言

近 20 年来,我国债券市场规模呈爆发式增长,但由于国家经济的整体情况呈现周期性回落的趋势,国家和市场经济结构加速调整,市场金融去杠杆的呼声日益增高,中国国内债券市场局部风险逐渐显现,债券违约率也水涨船高。2014 年 3 月份,出现了我国信用债市场的首例违约公募债——“11 超日债”,其出现打破了传统的“刚性兑付”思维[1]。随后爆发了一系列债券违约事件,直到 2019 年,我国市场债券违约的节奏才有所放缓,但是违约事件依然屡见不鲜。因此,建立债券违约预警模型对投资者加强风险防范和我国债券市场与国民经济稳定健康发展具有重要的现实意义。

为了能够将可能发生违约的债券主体提前识别出来,需要建立相关的债券违约预警模型。从学术研究角度,债券违约模型本质上是一个二分类问题,即在何种条件下债券可能发生违约和未违约。在二分类问题研究中,最常使用的是诸如神经网络、随机森林、logistic 回归以及支持向量机等经典可监督分类模型,这些模型可对样本中未加识别的债券进行违约预测判断。在实际问题的研究中,非均衡样本的问题大量存在。所谓非均衡样本是不同分类中一种分类占比过大,另一种分类占比过小,如银行信用检测、机械精度检测、医疗检测、语音信号处理、风险预警、信息检索等[2]。由于企业发生债券违约的数量在发行的信用债中所占比例非常低,所以样本非均衡化问题一直是债券违约研究的难点及重点。张永东[3]指出在分析不均衡样本的情况下,直接进行分类会导致分类器过多地关注数据中的多数类别,而忽视少数类别,进而使得分类结果具有明显的偏向多数类;Sun 等[4]研究也表明通常分类器将重点放在多的数类别上,因为它在样本中所占有的权重

收稿日期:2020-05-21;**修回日期**:2020-06-20.
作者简介:程建华(1964—),安徽宣城人,教授,从事经济预测、数据分析研究.
* **通讯作者**:徐恒宇(1995—),女,安徽安庆人,硕士研究生,从事金融风险管理研究. Email:1543859065@ qq. com.

比例较高。为了降低损失函数,分类器在整个数据集中识别多数类的样本可以达到很高的准确性,而对少数类样本的分类识别正确率自然会降低。当所研究的数据集高度不平衡时,算法的分类性能将受到重大影响。支持向量机(SVM)在非均衡样本下的应用研究主要包括改进算法和惩罚函数的权重以及平衡数据的正负类样本。而对于数据集层面的非均衡样本处理方法通常是人工平衡样本,主要是运用欠采样或者过采样的方法对训练集样本进行重构,从而降低样本的非均衡程度,以此来提高 SVM 分类器的预测准确率。

在经济领域,关于风险预警研究方面普遍存在非均衡样本问题,但国内学者较少将计算机领域的样本均衡化方法引入风险预警领域。令人可喜的是,近些年来国内学者开始重视风险预警方面的研究。张永东将 ADASYN 与 Logistic 相结合,通过过采样的方法改进 Logistic 的预测精度,在债券违约预警研究中,取得了良好效果;付君实运用不同的非均衡样本处理方法对传统的 SVM 模型进行改进,将 Borderline-SMOTE-Easy-Ensemble-SVM 模型引入极端金融风险预测中,预测精度显著提高。上述文献研究表明:样本均衡化方法较少应用于经济领域,非均衡样本问题也未在债券市场中引起重视。同样,在债券预警领域研究中,指标数量多且各个指标存在较强的相关性,即存在数据冗余问题。刘依恋[5]指出分类器会因数据维数过多而造成效率下降与过度拟合。目前,鲜有文献认识到冗余问题对于预警模型分类性能具有同样重要的作用。

基于以上研究,本文从数据层面入手,一方面引入样本数据均衡化算法,使用 ADASYN 算法自适应合成少数类样本,使得 SVM 训练的数据主要由多数类样本和均衡化的少数类样本构成;另一方面为进一步提高 SVM 预警模型的分类性能,将特征选择中的 mRMR 算法引入债券违约预警领域中,除去冗余、噪音数据,进而构建 ADmR-AdaBoostSVM 模型,选择预测效果最佳的参数。

1　研究方法

样本均衡化一直是数据挖掘研究中数据处理最为棘手的问题之一。面对数据不均衡的问题,SVM 同样存在误判少数类别的分类,而在债券违约预警研究中,违约样本应作为重点关注的样本,但对此不均衡样本数据的分类,单纯的 SVM 方法显得势单力薄。目前,针对不均衡数据,利用 SVM 分类存在两方面问题,一是算法问题,二是数据本身问题。算法方面的问题是指直接构建模型处理分类问题,如使用更高的权重损失用于少数类上,从而使得模型对于少数类别更为敏感;而数据方面处理方法则是利用适当的方法重构样本,使得数据的分布更加均衡,以提高分类器的性能。本文为提高 SVM 模型分类的准确性与有效性,对数据与模型两方面所存在的问题开展研究。针对数据不均衡问题,柳培忠等[6]研究发现 ADASYN 算法可以根据样本的分布情况来进行过采样,从而能够有效提高少类样本在边界区域的比例,能够缓解边界区域分布不平衡的问题,提高分类器的敏感度。下面是 ADASYN 算法的详细情况。

1.1　ADASYN 采样

ADASYN 方法属于自适应的合成采样算法,该算法主要思想是通过数据自身的分布情况来为少数类的样本自动生成新样本。

算法流程:

输入　假设训练集 D_{tr} 中具有 m 个样本为 $\{x_i, y_i\}$,$i=1,\cdots,m$,其中 x_i 是 n 维空间中的一个样本,$y_i \in Y=\{1,-1\}$ 是与 x_i 相关联类标签。定义 m_s 和 m_l 分别代表少数类和多数类样本数目,即 $m_s \leqslant m_l$,$m_s+m_l=m$。

① 计算不平衡度 $d=m_s/m_l$,$d \in (0,1]$。

② 计算合成样本量 $d=(m_s-m_l)\times\beta$。

③ 使用 k 近邻原则选择少数类样本的 k 个最近样本,使用欧氏距离度量近邻程度,记△为 k 个样本中多数类样本,记比例 r 为 $r_i=\Delta_i/K$,$i=1,\cdots,m_s$,$r_i \in [0,1]$。

④ 对于上述中的每一个由少数类计算的样本的 r_i,用 $\hat{r}_i = r_i / \sum_{i=1}^{m_s} r_i$,$r_i$ 为各比例的分布概率,其中 $\sum_i \hat{r}_i = 1$,得出每个少数类近邻的多数类比值的情况。

⑤ 根据④计算每个少数类样本需要合成的样本量 $g_i=\hat{r}_i \times G$。

⑥ 在每一个少数类样本的周围 k 个邻居中选择一个属于少数类的样本,然后由等式 $s_i=x_i+(x_{zi}-x_i)\times\lambda$ 进行合成数目为止。

1.2 最大相关最小冗余

由于数据信息行业的迅速发展,数据量的获取也变得容易,从而造成了数据维度的扩大。虽然在一定程度内,数据的分类准确率会随着维度的增长呈指数增长,但是在进行数据分析时,数据维度增加,解释变量过多还还会产生负面影响,比如线性回归模型多重共线性、过度拟合等问题。特别是当数据的维度过高、指标过多时,其中包含了过多内部相关项、冗余项和随机干扰项等,分类器会因数据维数过多而造成效率下降与过度拟合。高维数据的分类算法不仅使得模型精度下降,还会造成过拟合的风险,并由此带来"维数灾难"[7]。剔除最大相关项、降低数据维数、提高分类精度成为处理高维数据的主要方法之一。特征选择是通过从多个指标中选择少数最具有代表性的指标,用选择好的指标进行建模,能够使得与目标指标之间的信息量几乎完全保留,而各指标之间的信息量冗余较小。总的来说,特征选择可以通过降低特征维数,提升学习模型的训练速度从而达到比较好的训练效果[8]。特征提取和特征选择是目前主要的降维方式。特征提取是通过将某些原始特征或指标映射到更低维的空间,从而生成一些新的特性;而特征选择目的是找出原特征指标中最具有代表性的特征子集。特征选择不仅大幅度降低了数据维度,提高了分类器学习效率,而且可以提高分类器对各特征信息的学习功能,有效缓解过拟合现象[9]。

由于我国债券违约指标为公司的各个指标,所以数量较多且各个指标之间存在较强相关关系,必须要考虑数据冗余问题,并且其中存在着非线性关系,因此引入基于交互信息的提取特征方法:最大相关最小冗余(*m*RMR)算法。*m*RMR 的思想是利用交互信息量为参考指标来选择目标特征,通过惩罚已选取的各指标之间的冗余性使得选择的指标之间的相关性较小。

待选取的特征集 S 和目标类 c 的交互信息量由各个已选取特征 f_i 和用于分类的目标类 c 之间的所有交互信息量的平均值来决定,由式(1)定义:

$$D(S,c)=\frac{1}{|S|}\sum_{f_i\in S}I(f_i;c) \tag{1}$$

集合 S 中冗余信息定义用已选特征 f_i, f_j 之间的互信息值的平均值,如式(2)表示:

$$R(S)=\frac{1}{|S|^2}\sum_{f_i,f_j\in S}I(f_i;f_j) \tag{2}$$

*m*RMR 是结合上述两种定义,最大化式(3):

$$m\mathrm{RMR}=\max_S\left[\frac{1}{|S|}\sum_{f_i\in S}I(f_i;c)-\frac{1}{|S|^2}\sum_{f_i,f_j\in S}I(f_i;f_j)\right] \tag{3}$$

*m*RMR 算法是最大化待选特征指标与目标指标之间的信息量,最小化待选特征指标内部的冗余信息,指标的选择数目需进行定义。

1.3 Adaboost 算法

SVM 对不均衡数据中的多数样本非常敏感,此时分类器无法学习少数类样本特征,对于 SVM 可以选择多个核函数,从而能够提高该分类器的准确性,因此选择 Adaboost 以改进传统的 SVM 模型。AdaboostSVM 算法只需通过改变模型参数来拓展 SVM 分类器的精度范围。下面将详细描述 Adaboost 算法。

Adaboost 算法是 Yoav Freund 和 Robert Schapire 在 1997 年提出的解决分类的一种算法[8]。它采用对训练样本进行重新加权的方式产生不同的样本分布,中心思想是增加(减少)被错误(正确)分类的样本权重。开始时对于每一个样本设置权重,一般选择简单的平均值,在上一次分错的分类器样本的权重增大,其余的相应减少,对于更新权重后的样本继续重复上述步骤。对于每轮训练的结果,用总体样本再次训练弱分类器,然后赋予新的样本权值以及该弱分类器的权重,迭代至训练集完全正确或者实现决定的次数为止。

具体过程如下:给定样本 $x_i\in X$,分类 $y_i\in Y=\{0,1\}$,初始化 $D^1(i)=1/M$。当 $t=1,\cdots,T$ 时,有 $\{(x_1,y_1),\cdots,(x_M,y_M)\}$。

① 使用分布概率 D^t 训练基分类器 $h_t:X\to Y$。

② 计算误差:$\varepsilon_t=Pr_{i\sim D_i}(h_i(x_i)\neq y_i)$。

③ 选择权重更新参数 α_t。

④ 更新且归一化样本权重:$D^{t+1}(i)=Z_t^{-1}D^t(i)\exp(-\alpha_t h_t(x_i)y_i)$,其中 Z_t 为归一化因子。输出最终分类器,如式(4)所示:

$$H(x)=\mathrm{sign}\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right) \tag{4}$$

2 AdaboostSVM 模型

由于国内外经济金融形势的变化,为了社会融资成本的考虑,需要债券市场的稳定发展,一旦发生

债券违约,不但给上市公司带来重大损失,市场信用还将蒙受伤害,因此对债券违约进行预警和风险评估有助于促进证券市场稳定发展。本文在前人债券违约预警研究基础上,利用 AdaboostSVM 算法进一步开展债券违约预警研究,旨在提高债券违约预测的准确率。

AdaboostSVM 算法只通过改变参数来拓展成员 Adaboost 分类器的多样性。基于上述考虑,本文提出了适用于不平衡数据的分类模型 ADmR-AdaboostSVM 模型。该模型首先用 ADASYN 方法进行对少数类过采样,提高边界区域的少类样本比例,然后再利用 *m*RMR 进行特征选择,解决高维数据对分类器带来的"维数灾难",最后借助新的训练集对参数进行训练,得到最终的决策模型。

设训练数据集 D_{tr} 中具有 m 个样本为 $\{x_i, y_i\}$,$i=1,2,\cdots,m$,给定 $x_i \in X$,其中 $i=1,2,\cdots,16$,$y_i \in Y=\{0,1\}$,在进行 ADASYN 采样之后,平衡度 d 为 1;然后利用特征选择 *m*RMR 的方法对向量 $\boldsymbol{X}$ 求各个指标之间的最大相关和最小冗余,最后不断地训练模型使得参数达到最优。

建模步骤流程如图 1 所示:

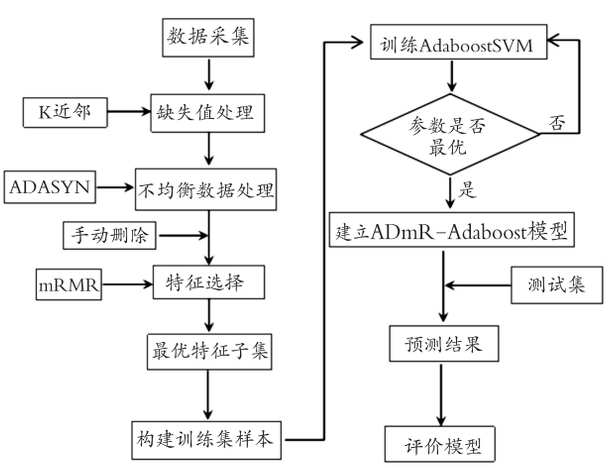

图 1　建模流程图

Fig. 1　Modeling flow chart

3　实证分析

3.1　样本的选取与数据来源

本文选取 2018 年上海交易所和深证交易所发行的信用债券(公司债与企业债)作为研究样本,其中交易所发行的信用债共有 468 只,实质违约的信用债有 30 只。

3.2　指标特征分析

蒋书彬[10]、蒋恒和杜立辉[11]对债券违约都做过相应研究,通过他们的研究发现,债券违约具有一定的阶段性以及过程性,具体表现为违约之前大多数企业会呈现财务危机过度、负债等问题。而过度负债是指企业付息能力不足、资产负债率过高和短期融资周转过慢,能够部分体现出企业经营不善[12]。而过度的无法偿还的债务使得发行人陷入困境,这会直接导致债务违约。过度负债、抗风险能力弱等是财务出现困境的主要特征,是债券违约产生的主要因素。

企业的正常经营多数情况下都需要负债,但债务需要通过公司对资产创造的收入、利润、现金流等来进行偿还,因此,企业的负债需要掌握好程度,如果企业的债务超过某个阈值,企业债券违约发生的概率将会增大。本文借鉴前人指标研究,从企业的资本结构、盈利能力、现金流量、偿债能力 4 个方面入手,准确地把握企业的财务危机,对信用债券违约进行预警研究。基于以上 4 个方面,筛选如下的 16 个指标进行后续分析,如表 1 所示。

表 1　预警指标体系

Table 1　Early warning index system

评估因素	候选指标变量
资本结构指标	股东权益率 X_1
	有形净值债务率 X_2
	长期负载率 X_3
	有息负债率 X_4
盈利能力指标	净资产收益率 X_5
	净利润与营业总收入之比 X_6
	资本收益率 X_7
	成本费用利润率 X_8
现金流量指标	现金流量与当期债务比 X_9
	债务保障率 X_{10}
	经营活动现金流量净额/净利润 X_{11}
偿债能力指标	流动比率 X_{12}
	速动比率 X_{13}
	资产负债率 X_{14}
	股权比例 X_{15}
	利润总额/债务总额 X_{16}

在建模之前,需要对数据进行筛选与清洗,其中包括脏数据清洗、缺失数据弥补与异常值的判定。对于缺失数据采取删除或者替补的方法进行处理。若对于某个样本存在大量缺失的指标无法提供有效信息,则删除这类指标。若存在个别财务指标缺失,

选择与其他指标相似度较高的指标,采取 K 近邻的方法进行替补。

3.3 样本均衡化

由于违约样本数据量明显少于非违约样本数据量,因此在建模之前采用 ADASYN 算法将数据进行扩充。违约样本为正向数据,其中训练集有 20 个正向数据和 292 个负向数据,本文使用 ADASYN 算法扩充了 271 个正向数据,这样数据之间基本达到均衡,从而适合 SVM 进行更方便的处理。

如图 2 和图 3 的二维图形是将成本费用利润率 X_8 和股权比例 X_{15} 两个指标进行可视化 ADASYN 操作的过程,其中 x 表示违约样本,o 表示非违约样本。经过过采样后样本数据达到均衡,而且从图中可以看出:总体样本中,违约样本几乎都在成本费用利用率左侧,这表明成本费用利用率越高越不容易违约,而股权比例表示未持有的股份比例,这说明未持有股份越高,越有可能导致违约风险提高,与经济学的规律相吻合,而在 ADASYN 抽样之后并没有消除这种隐含的关系,过抽样后数据的分布更明显,更有利于机器的学习。

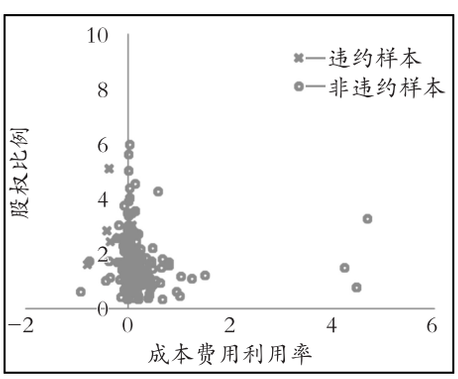

图 2 原始数据

Fig. 2 Raw data

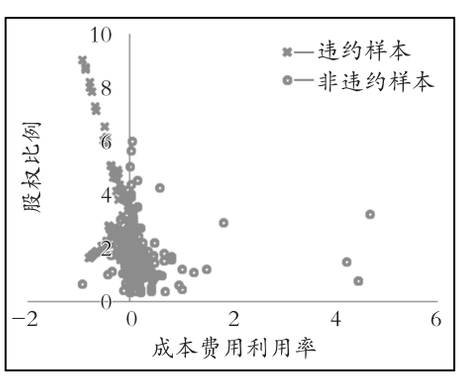

图 3 采样后数据

Fig. 3 Data after sampling

3.4 *m*RMR 特征提取

本文一共选取了 16 个指标,并将这些指标分为 4 个大类,见表 1,其中每个指标都是描述和反映公司企业生产、经营与管理等各个方面能力,不同指标之间存在不可避免的相关性与信息的交互性,而多余的信息对于本文所使用的 SVM 模型可能会起到过拟合的负面影响,所以对拟选的指标进行遴选是非常必要的。指标遴选有主成分分析(PCA)法和广义加性模型(GAM)方法等,而 PCA 只是对数据内部进行线性重组,得出较少的能够概括大部分信息的变量,并没有考虑到解释变量对目标变量的影响;GAM 方法则是利用非参数回归的置信度来表示解释变量对于目标变量的影响,没有考虑到各解释变量之间的交互信息冗余,以上两种方法都存在缺陷。实际上这 16 个特征之间存在着很强的相关性与非线性关系,为此,本文提出最大相关最小冗余方法用于减少冗余解释变量,从而降低数据分析的维度。与此同时,为了解释变量对目标变量的信息贡献度与解释变量内部交互信息冗余,本文综合以上两种方法的优点,而且保留各个变量的数据,使得对最后分类解释更加直观。

综上所述,本文研究基于债券违约数据,采用最大相关最小冗余算法(*m*RMR)进行特征的提取[13],将选择出的最优自己作为 SVM 模型的输入。在使用 *m*RMR 方法后,选取 4 个指标:长期负债率(X_3)、资本收益率(X_7)、成本费用利润率(X_8)以及股权比例(X_{15})。其中长期负债率为资本结构指标,它表示的是公司长期的负债状况,长期负债率越高,表示企业债务的负担越重,对企业的偿债能力产生负向的影响;资本收益率越高表面企业的预计盈利能力越强,能够带给公司充足的现金流,对企业的债务偿还与再融资帮助很大,对企业的债券偿还有着正向的作用;成本费用利润率是指企业一定期间的利润总额与成本、费用总额的比率,该指标衡量的是企业的全部劳动带来多少利润,能够综合反映出该企业的经济效应水平,能够代表企业的盈利质量,也能有一部分对企业的偿债有正向推动作用;股权比例和长期负债率都是表示企业的债务负担量,都是负向的影响。用这 4 个指标进行分类,理论上可以得到较好的结果,并在结果分析中得到验证。

3.5 模型评估

在评估不均衡样本数据集的分类性能时,传统

的性能评估指标已经不合适。针对传统性能评估存在的缺陷,采用二分类样本集的混淆矩阵,混淆矩阵会对预测结果与实际结果进行分类、对比、汇总,将样本划分为真正类(TP)、真负类(FN)、假正类(FP)、假负类(TN),具体如表 2 说明。

表 2　性能评估混淆矩阵

Table 2　Performance evaluation confusion matrix

实际结果	预测	
	正类	负类
正类	TP	FN
负类	FP	TN

本文使用 3 种评估标准。

(1)总体精度(O_A):

$$O_A=\frac{TP+TN}{TP+TN+FN+TN}$$

(2)正类预测值定义为查准率,表示预测正确的样本所占的比例:

$$p_{recision}=\frac{TP}{TP+FP}$$

(3)真实正类率或者叫查全率,表示正确正类占所有预测样本的比例:

$$r_{ecall}=\frac{TP}{TP+FN}$$

3.6　结果分析

本文利用 ADmR-AbaboostSVM 算法对债券违约进行预警判断。为了对所采用的分类算法做出评估,针对这一分类方法的性能在基于相同检验集的基础上,将其与 AdaboostSVM,*m*RMR-AdaboostSVM,AD-AdaboostSVM 3 种算法进行比较,结果如表 3 所示。

表 3　评估结果表

Table 3　Evaluation result table

算　法	O_A	$p_{recision}$	r_{ecall}
AdaboostSVM	0.92	0	0
mRMR-AdaboostSVM	0.92	0	0
AD-AdaboostSVM	0.76	0.19	0.67
ADmR-AbaboostSVM	0.77	0.20	0.67

如表 3 所示,总体精度(O_A)最高达到 92%,是由于原始数据的正类数据(违约样本)过少导致 AdaboostSVM 对于正类数据几乎无敏感度,分类器无法学习违约样本的特征,将所有的样本全部预测为非违约样本,虽然结果总体精度显著,但没有应用价值。相对于之后的改进模型,虽然总体精度更低,但更容易找出违约样本。在研究债券预警问题中,识别出违约债券才是关键。从表 3 中可以看出后两种方法要优于不经过过采样的数据集,而在 *m*RMR 提取信息之后的样本中可以发现查准率略有提高,说明 *m*RMR 方法应用在债券预警上有更好效果。为了进一步评价模型精度,本文通过比较各个算法的 ROC 曲线以及得分值做出更客观的评价(图 4—图 7)。

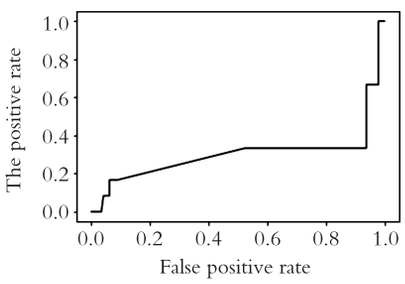

图 4　AdaboostSVM 模型 ROC 曲线

Fig. 4　ROC curve of AdaboostSVM model

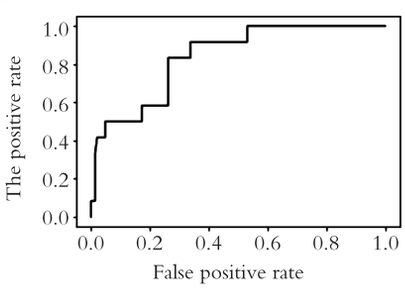

图 5　AD-AdaboostSVM 模型 ROC 曲线

Fig. 5　ROC curve of AD-AdaboostSVM model

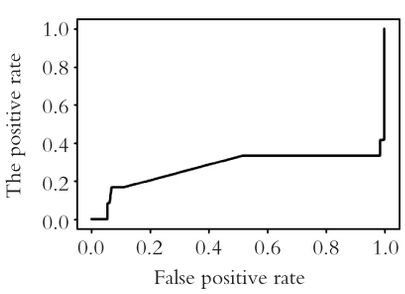

图 6　*m*R-AdaboostSVM 模型 ROC 曲线

Fig. 6　ROC curve of *m*R-AdaboostSVM model

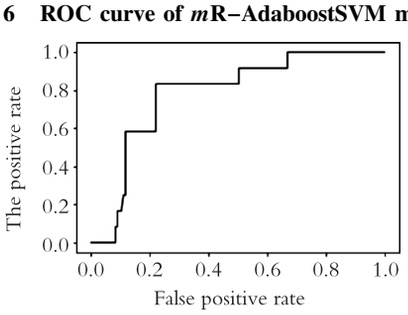

图 7　ADmR-AdaboostSVM 模型 ROC 曲线

Fig. 7　ROC curve of ADmR-AdaboostSVM model

表4 算法效果比较

Table 4 Comparison of algorithm effects

算 法	AUC	累计效能提升/%
AdaboostSVM	0.64	0
AD-AdaboostSVM	0.78	14
ADmR-AbaboostSVM	0.85	21

如图4所示,在非违约债券数量占比非常大时,ROC曲线很接近右下方,识别出违约债券的概率非常小,进行样本均衡化之后总体精度明显提高(如表4所示)。从图5和图6可知,在研究债券违约问题上,高维数据虽然会影响分类器的准确率,但远不及数据不均衡化给分类器带来误判率的灾难。由图7可知,在数据均衡化之后再进行特征选择,ROC曲线的面积进一步扩大,由表4可知,ADmR-AbaboostSVM的判别准确率在违约样本数量较小时可以达到85%左右,说明该模型获得了更好的分类性能。

4 结 论

本文选择Adaboost下的SVM模型是为了更好地拟合训练数据,单纯的SVM容易受样本量和缺失值的影响,而在Adaboost下可以很好地训练样本,使模型得到更好地拟合和预测能力。

债券样本中存在大量的不违约样本与少量的违约样本,如果直接运用常规的AdaboostSVM方法,则会使分类器对正常样本"过度学习",从而大大削弱分类器对违约的少数类样本的拟合能力和预测精度。本文通过信用债预警指标体系,从企业资本结构、盈利能力、现金流量、偿债能力4个维度刻画和评价发债主体的财务特征,使用ADASYN进行人工合成新样本,正负类配对比例为1∶1,改进了AdaboostSVM方法对违约样本预测性能,用检验集对模型进行预测,得出模型AUC值为78%,效能相较于AdaboostSVM模型提高了14%,取得了较好的预测效果。

特征选择是分类中的一个重要步骤,由于债券预警指标之间存在大量的相关性和冗余,通过利用*m*RMR特征选择的方法对变量进行特征选择,最终预测效果为85%,债券违约的识别率进一步提高。相较于样本均衡化,单纯的特征选择并未取得很好的效果,说明数据不均衡问题是影响债券预警模型精度的主要原因。

本文的不足之处在于债券公司的样本过少,无法使用交叉验证的方法来调整参数得到最优值,而只能通过以往的经验来确定参数的大小;其次应该更注重违约公司,对于很多无法分类为违约公司的确认违约公司,有待做出进一步分析。

参考文献(References):

[1] 彭兴韵.信用债券违约现状与对策[N].上海证券报,2016-05-27:12
PENG X Y. Current Situation and Countermeasures of Credit Bond Defaults[N]. Shanghai Securities News, 2016-05-27:12(in Chinese)

[2] 付君实.非均衡数据下基于SVM的极端金融风险预警研究[D].成都:成都理工大学,2016
FU J S. Research on SVM-based Extreme Financial Risk Early Warning Under Non-Equilibrium Data[D]. Chengdu: Chengdu University of Technology, 2016(in Chinese)

[3] 张永东. 基于非均衡样本的信用债违约风险预警研究[J]. 南方金融,2019(1):5—14
ZHANG Y D. Research on Early Warning of Credit Debt Default Risk Based on Unbalanced Samples[J]. Southern Finance, 2019(1):5—14(in Chinese)

[4] SUN Y, WONG A C, KAMEL M S. Classification of Imbalanced Data: A Review[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2009, 23(4):687—719

[5] 刘依恋.模式分类中特征选择算法研究[D].哈尔滨:哈尔滨理工大学,2014
LIU Y L. Research on Feature Selection Algorithm in Pattern Classification[D]. Harbin: Harbin University of Science and Technology, 2014(in Chinese)

[6] 柳培忠,洪铭,黄德天,等.基于ADASYN与AdaBoostSVM相结合的不均衡分类算法[J].北京工业大学学报,2017,43(3):368—375
LIU P Z, HONG M, HUNG D T, et al. Unbalanced Classification Algorithm Based on the Combination of ADASYN and AdaBoostSVM[J]. Journal of Beijing University of Technology, 2017, 43(3):368—375(in Chinese)

[7] XIE J Y, XIE W X. Several Selection Algorithms Based on the Discernibility of a Feature Subset and Support Vector Machines[J]. Chinese Journal of Computers, 2014,37(8):1704—1718

[8] 张俐,王枞. 基于最大相关最小冗余联合互信息的多标签特征选择算法[J]. 通信学报, 2018(5):111—122
ZHANG L, WANG Z. Multi-label Feature Selection Algorithm Based on Maximum Correlation and Minimum Redundancy Joint Mutual Information [J]. Journal of Communications, 2018(5):111—122(in Chinese)

[9] 杨文元. 基于最大相关最小冗余的多标记特征选择[J]. 数码设计,2016(2):21—25
YANG W Y. Multi-label Feature Selection Based on Maximum Correlation and Minimum Redundancy [J]. Digital Design, 2016(2):21—25 (in Chinese)

[10] 蒋书彬. 违约发债主体财务指标特征研究[J]. 债券, 2016(6):41—47
JIANG S B. Research on the Characteristics of Financial Indicators of Default Bond Issuers[J]. Bonds,2016(6): 41—47(in Chinese)

[11] 蒋恒,杜立辉. 我国债券市场违约成因分析及未来信用状况展望[J]. 债券,2016(5):43—47
JIANG H, DU L H. Analysis of the Causes of Default in China's Bond Market and Prospects for the Future Credit Status[J]. Bonds,2016(5):43—47(in Chinese)

[12] 龙章睿. 企业过度负债:涵义、识别与应对——基于银行信贷视角[J]. 南方金融,2016(6):54—62
LONG Z R. Excessive Corporate Debt: Meaning, Recognition and Response: Based on the Perspective of Bank Credit[J]. Southern Finance,2016(6):54—62(in Chinese)

[13] FEI S. A Hybrid Model of EMD and Multiple-kernel RVR Algorithm for Wind Speed Prediction [J]. International Journal of Electrical Power & Energy Systems,2016,78: 910—915

Research on Early Warning of Bonds Default Based on Unbalanced Data

CHENG Jian-hua, XU Heng-yu

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

Abstract:30 default bonds and 468 non-default samples were selected as research samples from Shanghai Stock Exchange and Shenzhen Stock Exchange, whether the bonds were defaulted was set as a binary classification problem for identification and analysis, and ADmR-AdaboostSVM classification model based on SVM was constructed for this problem. This article selects 16 early warning indicators from such four evaluation factors as enterprise capital structure, profitability, cash flow, and solvency, uses ADASYN method for oversampling and synthesizing new sample points, introduces mRMR method of feature extraction to bonds default field to obtain such four variables as long-term debt ratio, the rate of return on capital, profit margin on costs, and equity ratio as final early warning indicators of bonds default, and on this basis, uses AdaboostSVM model to conduct risk identification. Research results show that the sample unbalance problem is overcome during modeling process to make classification accuracy significantly improved, at the same time, the accuracy for identifying defaulted bonds is further achieved by solving the problem of high-dimensional data redundancy. Repeated verifications show that this model has strong robustness and effectiveness and has certain application value.

Key words:bonds default; ADASYN algorithm; mRMR algorithm; AdaboostSVM

责任编辑:李翠薇

引用本文/Cite this paper:

程建华,徐恒宇. 非均衡数据的债券违约预警研究[J]. 重庆工商大学学报(自然科学版),2021,38(3):86—93

CHENG J H, XU H Y. Research on Early Warning of Bonds Default Based on Unbalanced Data[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2021,38(3):86—93