文章编号：0253-374X(2022)02-0283-08

DOI：10.11908/j.issn.0253-374x.21106

基于不平衡数据的公司破产预测研究

周文泳[1]，冯丽霞[1]，段春艳[2]

(1. 同济大学 经济与管理学院，上海 200092；2. 同济大学 机械与能源工程学院，上海 201804)

摘要：整合创新数据预处理技术与集成算法利用不平衡数据探讨了公司破产预测问题。首先，运用冗余信息处理方法、不同抽样方法等对不平衡数据进行预处理。其次，以5.0分类器(Classifier 5.0，C5.0)决策树和单隐层前馈神经网络作为基分类器，分别与三类重抽样数据预处理技术结合，择出最优抽样法。再次，结合自助汇聚法提升分类效果，并运用十折交叉验证的受试者操作特征曲线的下方面积进行评价，对比了两基分类器的集成模型。最后，运用加利福尼亚大学尔湾分校数据库中一万多家波兰制造业公司的实际数据进行实验验证。实验结果表明：欠抽样或人工少数类过采样法与神经网络结合的集成模型分类效果最优，为企业实施破产预测提供积极支撑。

关键词：二元分类；不平衡数据；神经网络；C5.0决策树；集成方法

中图分类号：F272；TP183；O212　　**文献标志码**：A

Research on Company Bankruptcy Prediction Based on Unbalanced Data

ZHOU Wenyong[1], *FENG Lixia*[1], *DUAN Chunyan*[2]

(1. School of Economics and Management, Tongji University, Shanghai 200092, China; 2. School of Mechanical Engineering, Tongji University, Shanghai 201804, China)

Abstract: This paper discusses the problem of corporate bankruptcy prediction using unbalanced data by innovatively integrating data preprocessing technology and integration algorithm. Firstly, redundant information processing and different sampling methods are used to preprocess unbalanced data. Secondly, a decision tree with Classifier 5.0 (C5.0) and a single hidden layer feedforward neural network are used as the base classifier to select the optimal sampling method by combining with three kinds of resampling data preprocessing technologies. Thirdly, the self-aggregation method is combined to improve the classification performance, and the integration models of the two base classifiers are compared by the area under the receiver operating characteristic curve with 10-fold cross-validation. Finally, the actual data of more than 10 000 Polish manufacturing companies in the database of University of California Irvine are used for experimental verification. The experimental results show that the integrated model combining under-sampling or synthetic minority over-sampling method with neural network archive the best classification performance, which provides positive support for the enterprises to implement bankruptcy prediction.

Key words: binary classification; unbalanced data; neural network; C 5.0 decision tree; integrated methods

企业破产是企业财务困境中最为严峻的情形，企业经营状况不仅关系到企业的生存和发展，还影响到全球的经济，因此准确预测企业经营状况至关重要。传统的企业经营风险预测常常是决策者依据经验对企业当前情况进行判断，然而这对决策者提出了很高的要求，且决策过程也易受决策者的主观意识或外界因素干扰。如今随着大数据时代的到来，这种传统的预测方法已不能满足现代社会经济发展的需求。

早期建立了单变量判别模型[1]、多元线性判别模型[2]等用于破产预测的数学模型，而后多元逻辑回归模型在财务困境预测研究中渐渐发展，解决了判别分析中的许多问题，如受假设条件的约束[3]。自20世纪90年代以来，随着人工智能和机器学习的兴起，决策树、支持向量机、神经网络等先进技术在破产预测领域得到了快速发展，众多研究也证实了

收稿日期：2021-03-21
基金项目：2020年度同济大学“双带头人”教师党支部书记学术能力提升计划项目；上海市浦江人才计划(20PJ1413700)
第一作者：周文泳(1969—)，男，教授，管理学博士，主要研究方向为创新与技术管理、管理理论与工业工程。
E-mail：zhouwyk@126.com
通信作者：段春艳(1987—)，女，讲师，管理学博士，主要研究方向为工业工程、管理理论与工业工程等。
E-mail：duanchunyan77@163.com

论文拓展介绍

神经网络、决策树等机器学习算法在破产预测中具有更优的预测效果[4-9]。然而,实际预测的样本中往往是破产企业数量远小于未破产企业数量,样本数据的不均衡总是导致机器学习的预测性能下降。这一问题的主要特征表现为,在少数类样本量极少的情况下,分类器无法充分学习到少数类样本的特征,进而难以识别少数类样本。常见的解决思路是在数据层面将数据进行预处理,通过重抽样调整多数类与少数类的数量以实现类间样本量的平衡;此外在算法层面,运用集成学习算法对分类器进行增强[10]。Galar等[11]根据不同的基本集成学习算法和处理类不平衡问题的手段,划分了四类集成解决方案——代价敏感提升和数据预处理后分别基于提升、自助汇聚,以及结合提升与自举汇聚的双集成学习,并选择了4.5分类器(Classifier 4.5,C4.5)决策树作为基分类器,证明了在数据不平衡情形下,通过联合预处理技术(随机欠抽样等)和集成学习算法,可以获得更好的预测效果。而后也有一些研究基于支持向量机、人工神经网络、C4.5决策树等模型,将人工少数类过采样法(SMOTE)和自助汇聚、自适应提升等集成技术结合,获得了较好的分类结果[12-14]。Shen等[15]基于SMOTE抽样,对比了支持向量机、决策树等多种集成分类器,发现RF的分类效果较优。然而,过大的数据量会限制支持向量机的使用能力,此外决定其预测能力的核函数往往也需要慎重地手动选择[16]。相反,神经网络不仅适用于大样本,其自动提取数据特征的能力可一定程度上缓解核函数带来的问题[17]。目前已有学者将神经网络集成用于信息安全[18-19]、环境质量鉴别[20]、工业故障诊断[21]等多个研究领域,而用于公司破产预测领域的研究还较少。

因此,本文在前人研究基础上,选取神经网络和决策树作为基分类器,将数据预处理与集成算法结合构建公司破产预测模型,并对加利福尼亚大学尔湾分校(University of California Irvine,UCI)机器学习数据库提供的2007~2013年间一万多家波兰制造业公司进行实验。主要贡献包括:①在数据层面,选择三种重抽样方法——随机欠抽样、随机过抽样、SMOTE抽样进行预处理以实现类间样本量的平衡,并择优选出适合不同基分类器的抽样方法;②在算法层面,整合集成学习自助汇聚思想以提高单一分类器的预测效果。实验得出以神经网络为基分类器的模型结果优于以决策树为基分类器的模型结果,表明本文的研究方法能更有效地消除实际应用中不平衡数据的影响,且在企业破产预测领域具有较高的适用性,可为企业经营检测提供积极支撑。

1 研究方法

1.1 数据预处理技术

数据的预处理旨在预先对初始数据采取相关的审查、筛选、排序等必要措施[22]。数据预处理技术包含缺失、冗余信息处理,指标集优化筛选,标准化处理,抽样消除样本数据不平衡等多个阶段。

首先,初始样本数据往往存在缺失值,在所有待考察的属性下并非均有对应的数值,若不预先处理掉缺失值,会致使一些分类模型无法建立,如神经网络等。一般可通过特殊值、均值或众数等数值进行插补,而当存在缺失值的个案在数据集里的占比很小时亦可采取直接剔除的手段。

其次,在众多经济指标中,各指标之间难免会有相关性,因而导致数据冗余。若将所有指标直接代入建立分类模型,不仅会拖慢分类器的运行速度,还容易降低分类精度和模型的可解释性,因此选择类似主成分分析这样的手段根据指标间的相关性进行线性重组,进而得到能表示原始指标信息的少数几个综合性指标。

此外,为了像神经网络这样的模型能够较好地运行,其输入数据需进行标准化处理以消除量纲的影响,常见的方法如零-均值标准化、最小-最大标准化等。零-均值法适用于当数据呈正态分布时,通过转化函数为$X^{*}=\frac{X-\mu}{\sigma}$将其化为标准正态分布,其中$\mu$为样本数据的均值,$\sigma$为样本数据的标准差。而当数据呈现非正态或均匀分布时,可对每一个输入的数值型向量x,减去x中的最小值再除以x中值的范围以此将数据化至0~1范围内,函数表达式为$\frac{x-\min(x)}{\max(x)-\min(x)}$。

最后,由于分类器对不均衡数据集的有偏性,即多数类样本容易识别而少数类样本识别困难。本文分别通过随机过抽样、随机欠抽样、SMOTE抽样处理来平衡数据集。随机欠抽样主要是对多数类观测数采取随机剔除的方式,使得数据集达到平衡,该方法在数据量很大时非常有效。随机过抽样以随机重复少数类观测的方式来增添样本数目。SMOTE抽样也称人工数据合成法,利用生成人工数据来消除不平衡现象,而不仅是重复原始观测值。该方法基于特征空间(而非数据空间)产生与少数类观测相似

的新数据，而相似性则通过欧氏距离得以衡量。

1.2　机器学习算法

决策树算法是用于建立预测模型的有监督学习算法，是一种以树形结构来建立模型的递归划分探索法[17]，结构示意图如图1所示。

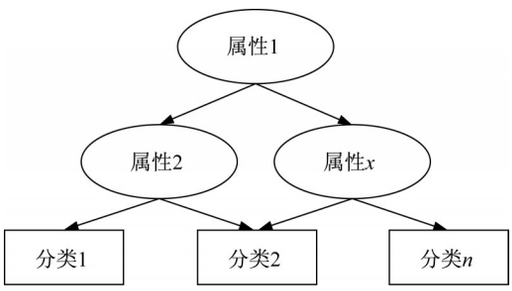

图1　决策树结构示意图

Fig.1　Schematic diagram of decision tree structure

决策树通过很多算法都可以实现，5.0分类器（Classifier 5.0，C5.0）是其最知名算法之一。它是由计算机科学家J. Ross Quinlan对之前的C4.5算法的改进，运算速度更快且更精准。C5.0决策树算法使用熵（entropy）来度量特征数据X的纯度，如式（1）所示[23]；然后再计算信息增益（gain）来决定根据哪一个特征进行分割，如式（2）所示[23]。决策树对于绝大多数的分类问题均适用。

$$\mathrm{Entropy}(X)=-\sum_{x\in X}p(x)\log p(x)\quad(1)$$

$$\mathrm{Gain}(X,Y)=\mathrm{Entropy}(X)-\mathrm{Entropy}(X|Y)\quad(2)$$

人工神经网络是通过仿照生物神经网络而开拓出来的进行信息处理的模型[24]。其中，多层前馈网络是应用最广泛和最受欢迎的人工神经网络之一，特别是在分类判别问题的应用中。图2显示了该网络的基本结构[25]。输入数据的特征数量直接决定网络输入层的节点个数，输出层的节点个数则由需要得出的结果数目决定。而对于隐藏层的节点个数，当下尚且并无一个绝对的标准。需要反复训练拥有不同节点数的模型，然后对比并适当地加或减其个数。隐层节点数目过大则使得模型易于出现过拟合，且计算量大、训练缓慢；过小则容易导致无法分类。

集成学习算法的核心是通过整合众多的单个弱学习器来建立强学习器。首先，输入训练数据建立多个模型，产生多个预测；之后，再利用投票表决或其他更复杂的方法来决定最终预测结果。使用集成学习的好处就是能节省寻找单一最佳模型的时间，并且由于集合了多个学习器的结果，也降低了单一学习器过拟合的可能性。自助汇聚法于1996年由

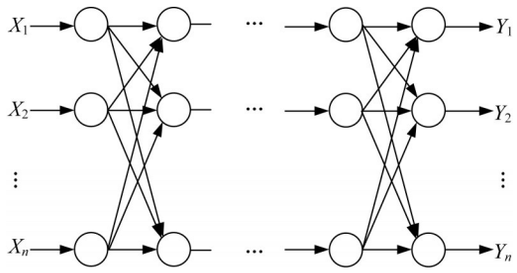

图2　多层前馈网络示意图[25]

Fig.2　Schematic diagram of multilayer feedforward network[25]

Breiman[26]提出，该集成方法通过在一个训练集合上重复训练进而得到多个分类器。它对相对不稳定的单一学习器（如决策树和神经网络（neural network，NN））能产生较好的分类效果，因为此类学习器会由于数据的细小改变而产生差别很大的模型。

1.3　基于不平衡数据的公司破产预测模型构建

本文将数据预处理技术与集成算法结合构建企业破产预测模型，在数据层面，涵盖缺失值处理、冗余信息处理、消除样本不平衡等多个阶段；再在算法层面，选取单隐层前馈神经网络和C5.0决策树作为基分类器，并结合集成学习自助汇聚思想来提高基分类器的预测性能。本文的模型构建路径如图3所示。

首先，将原始样本加载至R软件中进行初步的数据预处理。由于本文所用样本量较大，因此选择直接删除法进行缺失值处理。在冗余信息处理时，采用主成分分析法对通过缺失值处理的剩余指标进行降维，从纵向上精简输入属性的维数。然后按9∶1的比例将主成分分析之后产生的新数据集拆分为训练集和测试集。用0表示未破产类别，1表示破产类别。

其次，为了消除不均衡数据的影响，先在数据处理层面进行重抽样处理，分别通过随机过抽样、随机欠抽样、SMOTE抽样三种抽样处理，使两类数据量的比例达到1∶1。

进而，将随机过抽样、随机欠抽样、SMOTE抽样三种抽样技术分别与C5.0决策树、单隐层前馈神经网络两种基分类器相结合，创建6种不同的单一分类器，并在测试集上进行测试，通过比较选择出最适合各个基分类器的抽样方法。

最后，再从提高单一分类器性能的角度，将最优抽样技术与集成算法自助汇聚法结合，形成随机森林（random forest，RF）和神经网络集成两类集成分类器。其中，RF是通过C5.0决策树算法与自助汇

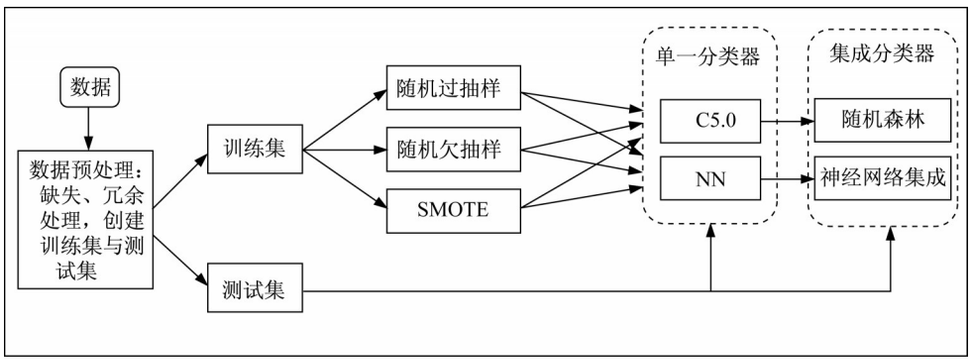

图3 基于不平衡数据的公司破产预测模型构建路径

Fig.3 Building path of corporate bankruptcy prediction model based on imbalanced data

聚算法整合，为决策树模型增添多样性；神经网络集成则是通过单隐层前馈神经网络与自助汇聚算法整合，以重新抽取训练数据集的方式来增添神经网络集成的差异程度。两者均是从横向角度对训练集实施多次选取得到多个有差异的网络个体，进而获得有差异的分类器。同样在测试集上进行检测，最终比较择出分类效果最佳的破产预测模型。

1.4 评价指标

(1)ROC曲线下面积

对于不平衡数据的分类问题，传统的分类精度评价准则确实能从宏观上映现分类性能，但并不表示能得出对的分类结果。因为当多数类样本数目远大于少数类时，后者能被识别的概率几乎为零。所以即使把所有样本都归为多数类，评价的精度依旧很高。Weiss等[27]的研究也证实，一般的分类精度评价标准会致使少数类的分类性能不佳。

受试者操作特征(receiver operating characteristic，ROC)曲线一般用于查验寻找真阳性与规避假阳性两者的权衡性。分别以假阳性比、真阳性比作为横、纵坐标画平面图，得到ROC曲线，示意图见图4[28]。为了更好地计量，计算ROC曲线的下方面积(area under the ROC，AUC)值来评判其二元分类的优劣，它表示预测的阳性类排在阴性类前面的概率。因其同时考虑了分类器对阳性类和阴性类的分类性能，因此即使在样本数据不平衡的情况下也能对分类器性能做出合理评价[29]。通常AUC的值使用如下评分体系：0.9～1.0=A(优秀)，0.8～0.9=B(良好)，0.7～0.8=C(一般)，0.6～0.7=D(较差)，0.5～0.6=F(无法区分)。

(2)十折交叉验证

本文所选取的决策树和神经网络两类基分类器均属于相对不稳定的学习器，为了使训练效果取得较为准确的评价，对每个模型都进行十折交叉验证。将数据集分成10部分，依次把9份合并当成训练集，剩余1份单独当成验证集来进行测验。每次试验都会产生相应的评价值，然后将10次结果的均值作为其最终评价。

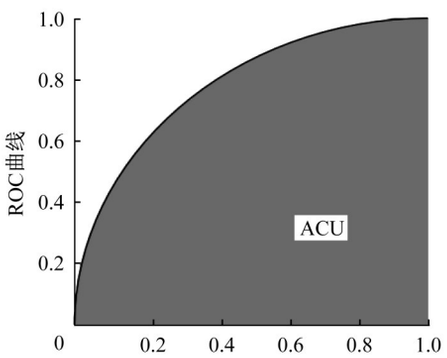

图4 ROC曲线[28]

Fig.4 The ROC curve[28]

2 研究设计

2.1 数据预处理

本文采用的波兰公司财务状况数据集由UCI机器学习数据库提供。样本数据包括64个财务指标，收集了近700家在2007－2013年间破产的公司和10 000多家仍在运营的公司数据。根据数据预测周期建立了5个分类案例。数据预处理步骤如下：

(1) 统一数据类型。将数据文件加载至R软件中，前64列的财务指标均转化为数值型，最后一列分类指标转化为因子型——“0”表示未破产，“1”表示破产。

(2) 缺失值处理。本文所用样本量较大，首先统计了有缺失数据的行，即指标数据有缺失的公司，发现超过50%的公司都有缺失数据。接着对列进行缺失值统计，发现了指标$x21$(销售(n)/销售(n-1))和$x37$((流动资产－存货)/长期负债)在5个预

测期的样本数据中存在着大面积的缺失，因此首先剔除掉这两个指标。此时再统计含有缺失值的公司个数，发现缺失率都降到了15%以内，这时即可直接删除这些公司数据。

(3)指标降维。利用主成分分析方法对通过缺失值处理的62个剩余经济指标进行降维。为消除各不同指标中量纲的影响，先将数据通过零－均值标准化，再用函数提取主成分。当方差累积贡献率至80%时即舍弃剩余的部分。本实验中第1至第5年的样本得到的主成分个数分别是10、10、13、15、14个，各碎石图如图5所示。

(4)创建随机的测试集和训练集。为保证分类器的训练效果，以9∶1的比例对主成分分析后的数据集进行划分，即90%的训练集和10%的测试集。划分后的样本数量如表1所示。

(5)抽样处理不平衡数据。对训练数据采取重抽样——分别通过随机过抽样、随机欠抽样、SMOTE抽样。利用R软件中的添加包ROSE(Random Over Sampling Examples，随机过抽样例子)以实现数据量1∶1的平衡。例如第1年有5 593个原始多数类样本，采用随机过采样法把少数类样本也增添至5 593个，由此数据集共有11 186条观测。

2.2　算法实现

本实验分别测试C5.0决策树模型和单隐层前馈神经网络模型在采用随机过抽样、随机欠抽样、SMOTE抽样技术后的分类性能，选择最适合本实验数据的抽样技术；然后再将最优抽样技术与自助汇聚法结合，寻找分类性能最佳的应用模型，并运用十折交叉验证得到的AUC值进行分类效果评价。

(1)三类抽样法与C5.0决策树。使用C5.0添加包建立决策树模型，将抽样技术处理过后的训练数据集用于训练C5.0决策树模型，然后再对仍旧保持不平衡状态的测试数据集进行测试。

(2)三类抽样法与神经网络。为了确保神经网络运行，其输入数据最好是在0附近，因此先将数据采用最小－最大标准化。使用NNET添加包构建单隐层前馈神经网络模型。对于参数的选择上，本文根据以往研究经验，对隐藏层的节点个数，依照经验公式$\sqrt{\text{输入节点数}\times\text{输出节点数}}$初始设置为5，再适当加上和减去一点余量，反复训练模型并测试[30]。其次为了更好防止过拟合，设置权重衰减参数。根据每个模型训练的实际情况进行权重衰减参数值在0.001～0.1之间的调整。

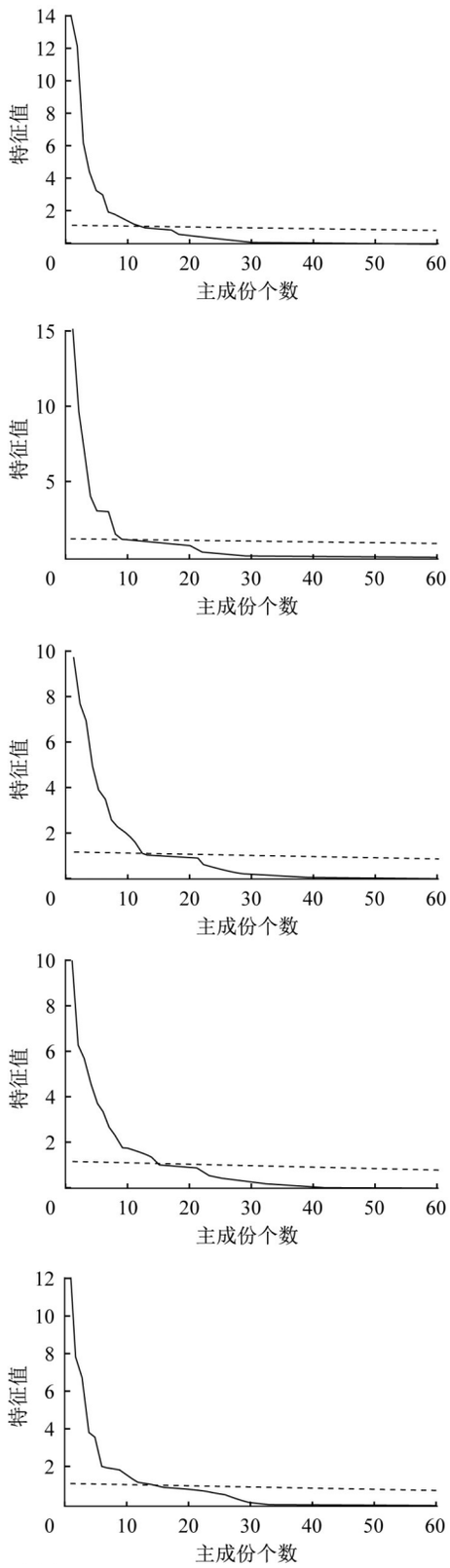

图5　第1至第5年的样本的主成分碎石图

Fig.5　Principal component lithotripsy diagrams of the samples from the first to the fifth years

(3)RF与神经网络集成。两者都是基于自助汇聚法思想，其中，RF是通过C5.0决策树算法与自助

表1 划分后的样本数量

Tab.1 The number of samples after partition

时间	训练集样本			测试集样本		
	未破产的	破产的	合计	未破产的	破产的	合计
第1年	5 593	93	5 686	620	12	632
第2年	7 520	197	7 717	836	22	858
第3年	7 685	268	7 953	857	27	884
第4年	7 077	293	7 370	800	19	819
第5年	4 335	228	4 563	482	25	507

汇聚法算法整合,为决策树模型增添多样性;神经网络集成则是将单隐层的前馈神经网络与自助汇聚法算法整合,以重新抽取训练数据集的方式来增添神经网络集成的差异程度。本文意图通过实验将两者进行比较。使用RF和CARET两个添加包分别构建RF模型和神经网络集成模型。最终所有模型都通过十折交叉验证求出AUC值,作为其分类评价结果。

3 结果分析

根据前面几节所介绍的步骤,把5个预测期的数据都分别代入单一分类器和集成分类器进行运算,得到对公司破产预测的分类评价结果如表2和表3所示。

(1)单一分类器比较

首先比较3种数据重抽样技术分别对C5.0决策树和单隐层前馈神经网络的分类效果的影响。对于C5.0决策树,通过五期AUC的平均值来比较三类不同抽样法,结果显示与欠抽样技术结合的决策树(C等)的分类性能更佳。因此为了后续模型性能的提升,选择欠抽样法与RF结合。对于单隐层前馈神经网络,发现与欠抽样、SMOTE抽样结合的模型性能都比较优良(均为C等),因此为后面阶段神经网络的集成选择随机欠抽样与SMOTE抽样。两种分类器的实验结果均显示欠抽样技术在处理不均衡数据上的良好效用。其次,比较两类单一分类器,从AUC平均数值上来看,欠抽样、SMOTE抽样下的单隐层前馈神经网络(0.79、0.77),更优于欠抽样下的C5.0决策树(0.73)。

表2 单一分类器的预测效果(AUC值及评价等级)

Tab.2 Prediction effect of single classifier (AUC value and evaluation grade)

分类模型	第1年	第2年	第3年	第4年	第5年	均值
过抽样+C5.0	0.53(F)	0.58(F)	0.68(D)	0.84(B)	0.83(B)	0.69(D)
欠抽样+C5.0	0.72(C)	0.73(C)	0.61(D)	0.75(C)	0.84(B)	**0.73(C)**
SMOTE抽样+C5.0	0.76(C)	0.54(F)	0.50(F)	0.70(C)	0.72(C)	0.64(D)
过抽样+NN	0.83(B)	0.53(F)	0.76(C)	0.80(B)	0.50(F)	0.68(D)
欠抽样+NN	0.83(B)	0.76(C)	0.74(C)	0.79(C)	0.83(B)	**0.79(C)**
SMOTE抽样+ NN	0.84(B)	0.79(C)	0.76(C)	0.79(C)	0.67(D)	**0.77(C)**

表3 集成分类器的预测效果(AUC值及评价等级)

Tab.3 Prediction effect of ensemble classifier (AUC value and evaluation level)

分类模型	第1年	第2年	第3年	第4年	第5年	均值
欠抽样+RF	0.75(C)	0.74(C)	0.74(C)	0.88(B)	0.81(B)	0.79(C)
欠抽样神经网络集成	0.84(B)	0.77(C)	0.74(C)	0.80(B)	0.83(B)	**0.80(B)**
SMOTE神经网络集成	0.84(B)	0.79(C)	0.76(C)	0.80(B)	0.83(B)	**0.80(B)**

(2)集成分类器比较

从集成学习的角度比较集成分类器和单一分类器,结果显示无论是通过集成学习之后的决策树模型还是集成神经网络,模型的预测性能都有所提升。尤其是针对那些集成前分类效果较差的预测期数据,模型集成后其性能有显著的提升,如第3年数据的决策树模型(从D到C)和第5年数据的神经网络模型(从D到B)。再从不同分类器的角度比较随机欠抽样下的RF,与随机欠抽样、SMOTE抽样下的神经网络集成,从评价均值上来看后两者对于公司破产预测效果更优(C、B、B)。

4 结论

2020年新冠疫情的爆发更是加大了企业对破产预测的重视程度。本文着眼于破产预测中样本类别数据不均衡且样本规模较大的问题,从增加分类器差异度的角度,对传统的预测模型进行改进,建立了基于重抽样技术和自助汇聚集成算法两者联合的机器学习模型,并对UCI机器学习数据库中一万余条波兰制造业公司数据进行实验。本研究选取C5.0决策树与单隐层前馈神经网络两种基分类器,结合数据层和算法层两方面的改进,并通过十折交叉验

证的AUC值进行评判。

最终实证结果显示：

（1）针对类别不平衡的公司破产预测样本，随机欠抽样和SMOTE抽样技术能辅助单一分类器获得更优良的预测效果；

（2）进而结合集成学习自助汇聚思想时，神经网络集成模型的预测结果不仅优于其单一分类器模型，也更优于RF模型。本文构建的预测模型结合了数据层面和算法层面的改进，通过大量的样本数据进行模型训练，有效消减了实际应用中不平衡训练集带来的影响，得到了具有较好预测性能的集成分类器，能准确预测公司破产风险，可应用于记录了众多财务指标属性的公司数据集中，为公司经营检测提供积极支撑，进而使公司及早实施相关措施预防破产。

总而言之，建立科学、准确且实用的公司破产预测模型，不仅能够帮助企业管理者及时地识别公司潜在的经营风险，还能帮助投资者等众多利益相关方做出正确的投融资决定以免遭受巨大损失，同时对国家及地方政府的资金、人力等投入规划的制定也具有重要的辅助作用。此外，随着技术的不断更新升级，公司破产预测模型也依旧是在不断变化中发展的。从起初企业家的经验判断、判别分析，到如今的机器学习算法，公司破产预测研究始终是一个永恒且热门的话题。面对新时代背景下不断涌现的新的难题与挑战，未来应不断探索新途径，持续对破产预测模型进行调整和创新，以完善公司破产预测领域的研究。

未来研究可进一步从此方向入手：针对神经网络这种黑箱方法，建立更优参数配置的神经网络集成模型，提升集成学习后的预测效果；除了常用的单隐层前馈神经网络，还可以尝试采用径向基神经网络等其他方法。

作者贡献声明：

周文泳：指导研究方案和论文撰写，全文审阅。

冯丽霞：数据收集与处理，算法实现，撰写论文。

段春艳：指导研究方案和论文撰写，审阅及修订论文。

参考文献：

[1] BEAVR W H. Financial ratios as predictors of failure, empirical research in accounting: selected studies[J]. Journal of Accounting Research, 1966(4):71.

[2] ALTMAN E I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy[J]. The Journal of Finance, 1968, 23(4):589.

[3] 周兴荣，李文宁. 国内外财务预警方法与模型评析[J]. 财会月刊，2008(15):79.
ZHOU Xingrong, LI Wenning. Financial early warning methods and models at home and abroad [J]. Financial Monthly, 2008(15):79.

[4] ODOM M, SHARDA R. A neural network model for bankruptcy prediction [C]// Proceedings of the IEEE International Conference on Neural Network. New York: IEEE, 1990:163-168.

[5] 蔡福安. 神经网络在银行破产预测中的应用[J]. 预测，1993，12(6):45.
CAI Fu 'an. Application of neural network in bank bankruptcy prediction[J]. Forecast, 1993, 12(6):45.

[6] 刘旻，罗慧. 上市公司财务危机预警分析——基于数据挖掘的研究[J]. 数理统计与管理，2004(3):51.
LIU Min, LUO Hui. Financial crisis early warning analysis of listed companies——research based on data mining [J]. Mathematical Statistics and Management, 2004(3):51.

[7] 姚靠华，蒋艳辉. 基于决策树的财务预警[J]. 系统工程，2005，23(10):102.
YAO Kaohua, JIANG Yanhui. Financial early warning based on decision tree[J]. Systems Engineering, 2005, 23(10):102.

[8] 吴俊杰. 财务困境预测：数据挖掘方法的比较与运用[J]. 清华大学学报(哲学社会科学版)，2006(S1):47.
WU Junjie. Financial dilemma prediction: comparison and application of data mining methods [J]. Journal of Qinghua University (Philosophy and Social Sciences Edition), 2006 (S1):47.

[9] 阎娟娟，孙红梅，刘金花. 支持向量机的上市公司财务危机预警模型[J]. 统计与决策，2006(12):158.
YAN Juanjuan, SUN Hongmei, LIU Jinhua. Financial crisis early warning model of listed companies based on support vector machine[J]. Statistics and Decision, 2006(12):158.

[10] 李艳霞，柴毅，胡友强，等. 不平衡数据分类方法综述[J]. 控制与决策，2019，34(4):673.
LI Yanxia, CHAI Yi, HU Youqiang, *et al*. A review of unbalanced data classification [J]. Control and Decision, 2019, 34(4):673.

[11] GALAR M. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches [J]. IEEE Transactions on Systems Man & Cybernetics Part C Applications & Reviews, 2012, 42(4):463.

[12] 商志明. 基于非平衡视角的企业财务困境智能预警研究[D]. 金华：浙江师范大学，2013.
SHANG Zhiming. Research on intelligent early warning of enterprise financial dilemma based on non-equilibrium perspective [D]. Jinhua: Zhejiang Normal University, 2013.

[13] SUN J, LI H, HAMIDO F, *et al*. Class-imbalanced dynamic financial distress prediction based on adaboost-svm ensemble combined with SMOTE and time weighting [J]. Information Fusion, 2020, 54:128.

[14] CHOI H, SON H, KIM C. Predicting financial distress of contractors in the construction industry using ensemble learning [J]. Expert Systems with Applications, 2018, 110: 1.

[15] SHEN F, LIU Y, WANG R, *et al*. A dynamic financial distress forecast model with multiple forecast results under unbalanced data environment [J]. Knowledge-based Systems, 2020, 192: 105365.

[16] 汪海燕,黎建辉,杨风雷.支持向量机理论及算法研究综述[J].计算机应用研究,2014,31(5):1281.
WANG Haiyan, LI Jianhui, YANG Fenglei. A review of support vector mechanism and algorithms [J]. Application Research of Computers, 2014, 31(5): 1281.

[17] LANTZ B,李洪成,许金炜,等.机器学习与R语言[M].北京:机械工业出版社,2015.
LANTZ B, LI Hongcheng, XU Jinwei, *et al*. Machine learning and R language [M]. Beijing: China Machine Industry Press, 2015.

[18] 翟璐.一种基于Boosting算法的新模型在银行信用评级中的应用[D].北京:北京交通大学,2016.
ZHAI Lu. Application of a new model based on boosting algorithm in bank credit rating [D]. Beijing: Beijing Jiaotong University, 2016.

[19] 傅康.基于CNN的信用卡欺诈检测[D].上海:上海交通大学,2017.
FU Kang. Credit card fraud detection based on CNN [D]. Shanghai: Shanghai Jiao Tong University, 2017.

[20] 蒋洪迅,田嘉,孙彩虹.面向PM_(2.5)预测的递归随机森林与多层神经网络集成模型[J].系统工程,2020,38(5):14.
JIANG Hongxun, TIAN Jia, SUN Caihong. A recursive random forest and multi-layer neural network integrated model for PM _(2.5) prediction [J]. Systems Engineering, 2020, 38(5): 14.

[21] 崔宇,侯慧娟,苏磊,等.考虑不平衡案例样本的电力变压器故障诊断方法[J].高电压技术,2020,46(1):33.
CUI Yu, HOU Huijuan, SU Lei, *et al*. Fault diagnosis method of power transformer considering unbalance case sample [J]. High Voltage Technology, 2020, 46(1): 33.

[22] 马冀,赵养森,罗宏.统计基础与实用方法[M].上海:立信会计出版社,2012.
MA Ji, ZHAO Yangsen, LUO Hong. Statistical basis and practical methods [M]. Shanghai: Lixin Accounting Publishing House, 2012.

[23] 赵丽,程铁信,莫莹,等.基于C5.0改进算法的焊接工艺参数选择决策树数据挖掘模型及其应用[J].中国管理科学,2016,24(S1):177.
ZHAO Li, CHENG Tiexin, MO Ying, *et al*. Data mining model of welding process parameters selection decision tree based on improved C5.0 algorithm and its application [J]. Chinese Management Science, 2016, 24(S1): 177.

[24] 杨剑锋,乔佩蕊,李永梅,等.机器学习分类问题及算法研究综述[J].统计与决策,2019,35(6):36.
YANG Jianfeng, QIAO Peirui, LI Yongmei, *et al*. A review of machine learning classification problems and algorithms [J]. Statistics and Decision, 2019, 35(6): 36.

[25] 江学军,唐焕文.前馈神经网络泛化性能力的系统分析[J].系统工程理论与实践,2000(8):36.
JIANG Xuejun, TANG Huanwen. Systematic analysis of generalization capability of feedforward neural networks [J]. Systems Engineering Theory and Practice, 2000(8): 36.

[26] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123.

[27] WEISS G M, Provost F J. Learning when training data are costly: the effect of class distribution on tree induction [J]. Journal of Artificial Intelligence Research, 2003, 19: 315.

[28] 周志华.机器学习[M].北京:清华大学出版社,2016.
ZHOU Zhihua. Machine learning [M]. Beijing: Tsinghua University Press, 2016.

[29] HE H, GARCIA E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263.

[30] 马冬梅.基于深度学习的图像检索研究[D].呼和浩特:内蒙古大学,2014.
MA Dongmei. Research on image retrieval based on deep learning [D]. Hohhot : Inner Mongolia University, 2014.