文章编号： 1673-3193（2024）03-0257-08

基于改进SMOTE算法和Ensemble模型的学习结果预测方法

王晓勇[1]，胡胜利[2]

（1. 淮南联合大学 信息工程学院，安徽 淮南 232038；2. 安徽理工大学 计算机科学与工程学院，安徽 淮南 232001）

摘　要： 为解决不同领域的数据分类和预测任务中单个机器学习算法适用性较差的问题，以及缓解数据集严重不平衡对预测性能的影响，提出了基于合成少数类过采样（SMOTE）和Ensemble集成模型的数据分类方法。传统SMOTE算法通过对少数类样本进行插值来生成新的合成样本，合成样本中存在噪声和样本间相似性较高的问题。为此，提出了改进的SMOTE算法，通过距离计算移除噪声样本和易混淆样本，得到高区分度的纯净合成样本。然后，利用Ensemble方法调整样本和分类器权重，并组成分类效果更好的强分类器。在公开在线学习数据集Kalboard360上的实验结果表明，使用极限随机树（ERT）分类器时，结合改进SMOTE和Ensemble模型后实现了97.9%的预测准确度，比单个ERT分类器提升了5.5%，证明所提改进SMOTE算法能够生成高质量的均衡化数据，且集成学习模型的性能显著优于单个机器学习算法。

关键词： 机器学习；神经网络；数据挖掘；集成学习；数据均衡化；学习结果预测

中图分类号： TP391　　**文献标识码：** A　　**doi：** 10.3969/j.issn.1673-3193.2024.03.002

引用格式： 王晓勇，胡胜利．基于改进SMOTE算法和Ensemble模型的学习结果预测方法[J]．中北大学学报（自然科学版），2024，45（3）：257-264.

WANG Xiaoyong，HU Shengli. Learning result prediction based on improved smote algorithm and ensemble model[J]. Journal of North University of China（Natural Science Edition），2024，45（3）：257-264.

Learning Result Prediction Based on Improved SMOTE Algorithm and Ensemble Model

WANG Xiaoyong[1]，HU Shengli[2]

（1. School of Information Engineering，Huainan Union University，Huainan 232038，China；
2. School of Computer Science and Engineering，Anhui University of Science Technology，Huainan 232001，China）

Abstract： In order to solve the problem of poor applicability of a single machine learning algorithm in data classification and prediction tasks in different fields，and to alleviate the impact of severe imbalance in datasets on prediction performance，a learning result prediction method based on Synthetic Minority Oversampling（SMOTE）and the ensemble model was proposed. The traditional SMOTE algorithm generated new synthetic samples by interpolating minority class samples，which could result in the presence of noise and high similarity between synthetic samples. To address these issues，an improved SMOTE algorithm was proposed，which removed noisy and easily confused samples by distance calculation，resulting in high discriminative and pure synthetic samples. Subsequently，an ensemble method was utilized to adjust the

收稿日期： 2023-04-11

基金项目： 安徽省重点科研项目（KJ2021A1306）

作者简介： 王晓勇（1978－），男，教授，硕士，主要从事深度学习、大数据分析等研究。

weights of samples and classifiers, leading to the creation of a stronger classifier with improved classification performance. Experimental results on the public online learning dataset Kalboard360 show that when using the Extreme Randomized Trees (ERT) classifier, in combination with improved SMOTE and Ensemble model, resulted in a prediction accuracy of 97.9%, which is a 5.5% increase compared to using a single ERT classifier. This demonstrates that the proposed SMOTE algorithm can generate high-quality balanced data, and the performance of the Ensemble learning model is significantly better than that of a single machine learning algorithm.

Key words: machine learning; neural networks; data mining; ensemble learning; data balancing; learning outcome prediction

0 引 言

当前，信息和通信技术的快速发展带来了海量的数据，传统统计分析方法无法满足大数据背景下知识发现的需求。数据挖掘通过使用复杂的统计工具集合或人工智能模型，从不同角度进行数据分析，提取数据内部关联，将累积的非结构化信息转换为直观数值，从而促进决策制定过程[1]。随着在线学习平台的兴起，利用各种机器学习算法对学习数据进行深度挖掘已成为研究热点[2]。

学习数据挖掘(Learning Data Mninng，LDM)通过分析学习过程并预测学习效果来改善知识的传递方式和吸收效果[3]。陈曦等[4]结合知识图谱表征和协同过滤算法对稀疏数据场景下的学习效果进行了预测，通过知识相似度计算和历史数据补足提高了预测性能。Trakunphutthirak等[5]认为单个数据源不足以识别出学习者的特征，提出了网页浏览和互联网活动数据集，对时域和频域中的不同权重进行了比较，并使用随机森林(Random Forest，RF)分类器实现了最好的预测性能。申航杰等[6]提出了结合模糊聚类和支持向量机回归的预测方法，其中考虑到了历史数据和行为习惯等因素。

神经网络已成LDM中较为成熟的分类和预测模型，与传统网络模型及其学习算法相比，其适用于处理大规模数据，且能够同时应用到结构化和非结构化数据中。林梦楠等[7]提出了基于BP神经网络的预测方法，并通过自适应差分算法进行权值优化，取得了较好的预测结果。张阳等[8]提出了结合图卷积网络(Graph Convolutional Network，GCN)和自编码器模型的预测方法，利用GCN进行编码并获取了数据内在特征，最后输入解码器得到预测结果矩阵。

近些年，Ensemble分类器已成为机器学习和模式识别研究中的流行策略。Ensemble是一种结合不同分类器输出的方法，通过对多个弱分类器进行加权并合并为单个强分类器来提高分类效率，从而实现比个体分类器更好的处理性能[9]。李慧芳等[10]提出了多维特征融合策略，基于Ensemble学习策略合并类别型、数值型和原始输入特征，提高了云工作流任务的预测精度。Kamal等[11]提出了基于Ensemble的树分类器，并利用统计工具分析了特征选择对预测性能的影响。

传统SMOTE算法通过对少数类样本进行插值来生成新的合成样本，其合成样本中存在噪声和样本间相似性较高的问题。为此，提出改进SMOTE算法，通过距离计算移除噪声样本和易混淆样本，得到高区分度的纯净合成样本。同时，利用Ensemble方法调整样本和分类器权重，组成分类效果更好的强分类器。本文的主要创新点总结如下：

1)提出了改进的合成少数类过采样(Synthetic Minority Oversampling Technique，SMOTE)算法，基于原始少数类样本与新生成少数类样本之间的距离进行调整，移除噪声样本和易混淆样本，解决了原算法中的噪声和边界样本的问题，在实现数据集平衡的同时确保了模型性能的稳定。

2)将Ensemble算法用于LDM任务中，使用不同机器学习模型作为基础分类器，在每轮训练中逐步增加误分类样本在基础分类器中的权重，将多个弱分类器合并为单个强分类器，提升了模型的预测性能。

1 本文提出的方法

图1给出了所提基于改进SMOTE算法和

Ensemble模型的学习效果预测方法的流程图，其中包括预处理、模型开发和性能评估阶段。在预处理阶段，首先将采集到的数据转换为合适的格式。通过离散处理将学习效果转换为与分类标签相对应的标称值。为解决分类不平衡对机器学习算法的影响，利用改进的SMOTE算法来改善训练数据的平衡度。然后，使用监督分类算法训练模型，学习从输入特征到输出标签的映射函数。

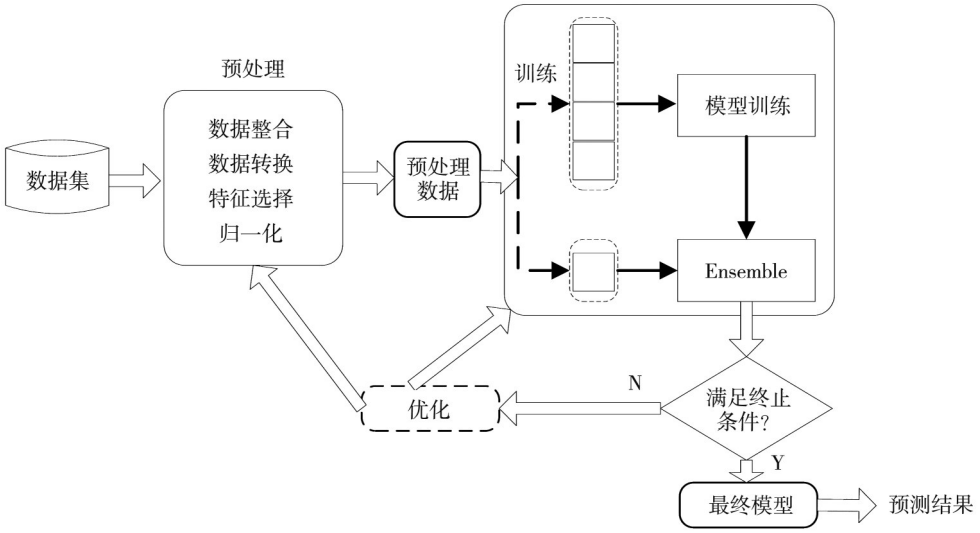


图 1　本文方法流程图

Fig. 1　Flowchart of the method in this paper

1.1　改进的SMOTE算法

在数据预处理阶段，使用改进的SMOTE处理训练数据中分类不平衡的问题。大部分分类算法旨在聚集纯净样本并从中学习，将每个类别的边界尽可能定义得更清晰，从而提高预测性能。远离分类边界的人工实例更易于分类，而靠近分类边界的人工实例会提高分类器学习的难度。

为更清晰地区分不同类别的边界线，在数据集平衡处理中生成更纯净的样本，提出了改进的SMOTE算法。首先，使用SMOTE创建人工实例[12]

$$N=2*(r-z)+z, \tag{1}$$

式中：r为多数类样本；z为少数类样本数量；N为新生成人工实例的数量。令$\hat{x}=\{\hat{x}_1,\hat{x}_2,\hat{x}_3,\cdots,\hat{x}_N\}$为新的人工实例集合，$\hat{x}_i^{(j)}$为$\hat{x}_i$的第$j$个属性值，$S_m=\{S_{m1},S_{m2},\cdots,S_{mz}\}$和$S_a=\{S_{a1},S_{a2},\cdots,S_{ar}\}$分别为少数类样本和多数类样本。计算$\hat{x}_i$与$S_{mk}$之间的距离

$$D_{\min}(\hat{x}_i,S_{mk})=\sum_{j=1}^{M}\sqrt{(\hat{x}_i^{(j)}-\hat{S}_{mk}^{(j)})^2}。 \tag{2}$$

然后，计算$\hat{x}_i$与之间S_{al}的距离

$$D_{\text{maj}}(\hat{x}_i,S_{al})=\sum_{j=1}^{M}\sqrt{(\hat{x}_i^{(j)}-\hat{S}_{al}^{(j)})^2}。 \tag{3}$$

由此，分别计算出数组$A_{\min}$和A_{maj}。

$$A_{\min}=(D_{\min}(\hat{x}_i,S_{m1}),\cdots,D_{\min}(\hat{x}_i,S_{mz})), \tag{4}$$

$$A_{\text{maj}}=(D_{\text{maj}}(\hat{x}_i,S_{a1}),\cdots,D_{\text{maj}}(\hat{x}_i,S_{ar}))。 \tag{5}$$

从$A_{\min}$和A_{maj}中分别选出最小值$\min(A_{\min})$和$\min(A_{\text{maj}})$。若$\min(A_{\min})<\min(A_{\text{maj}})$，则接受新生成的人工样本，否则，拒绝该样本。

令$\hat{S}=\{\hat{S}_1,\hat{S}_2,\cdots,\hat{S}_n\}$为已接受的人工样本，执行噪声移除步骤。计算$\hat{S}$与每个原始少数类样本$S_m$之间的距离

$$\min(\hat{S}_i,S_m)=\sum_{k=1}^{z}\sum_{j=1}^{M}\sqrt{(\hat{S}_i^{(j)}-S_{mk}^{(j)})^2}。 \tag{6}$$

将$\hat{S}$与每个原始多数类样本S_a之间的距离计算为

$$\text{maj}(\hat{S}_i,S_a)=\sum_{i=1}^{r}\sum_{j=1}^{M}\sqrt{(\hat{S}_i^{(j)}-S_{al}^{(j)})^2}。 \tag{7}$$

基于式(6)和式(7)计算出的距离指标，移除与原始多数类样本距离较小的半数易混淆的合成样本，由此得到高纯度的合成样本，从而改善训练数据集的平衡度。

1.2　机器学习算法

所提方法应用监督式分类算法，通过模型训练学习从输入特征到输出标签的映射。本文测试了4种不同的机器学习算法在LDM任务中的预测性能。

1.2.1 支持向量机

支持向量机(Support Vector Machine, SVM)是一种监督式学习算法,将实例表示为N维空间中的点集,并通过生成$(N-1)$维的超平面对数据点分组。SVM旨在发现模式x的超平面,并将线性决策边界最大化,即

$$\boldsymbol{y}(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0, \tag{8}$$

式中:$\boldsymbol{y}(\boldsymbol{x})=0$定义了特征空间中的决策边界;$\boldsymbol{w}$为法向量;$b$为偏置。

不同类别的数据之间的最优超平面可求解为二次优化问题

$$\min Q(\boldsymbol{w},b,\boldsymbol{\xi})=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}\boldsymbol{\xi}_i, \tag{9}$$

式中:$\forall_i y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}(x_i)+b)\geqslant 1-\boldsymbol{\xi}_i$,$\boldsymbol{\xi}_i\geqslant 0$,$\boldsymbol{\Phi}$为特征向量;$\boldsymbol{\xi}$为测量误分类误差的松弛变量;$\boldsymbol{C}$为控制变量。

本文使用了通过卡方统计进行特征选择的SVM算法[13]。对寻找最优超平面,利用核函数将训练数据自动转换到高维空间,从而使数据线性可分。将核函数应用到原始空间中的数据,定义为

$$K(x_i,x_j)=\boldsymbol{\Phi}(x_i)^{\mathrm{T}}\boldsymbol{\Phi}(x_j)。 \tag{10}$$

本文使用了高斯径向基函数作为SVM的核函数

$$K_{\mathrm{Gaussian}}(x_i,x_j)=\exp\left(\frac{\|x_i-x_j\|^2}{2\sigma^2}\right), \tag{11}$$

式中:σ为核参数,控制高斯函数的宽度。

1.2.2 贝叶斯网络

贝叶斯网络(Bayesian Network, BN)是得到广泛使用的概率图模型,能够促进不确定条件下的推理和预测。利用有向无环图进行推导,其中属性表示为节点,属性相关的假设则表示为弧线。BN分类器可定义为

$$c_{\mathrm{Bayes}}=\arg\max_{c\in C}P(c)P(x_1,x_2,\cdots,x_n|c), \tag{12}$$

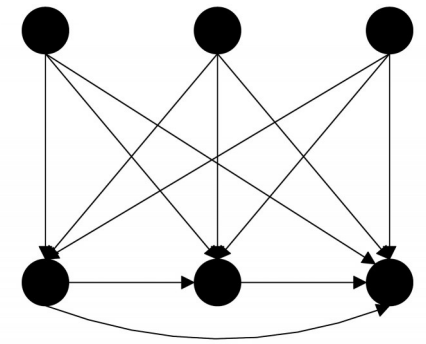
式中:x_i为属性值;c为类别标签。

最简单的BN分类器为朴素贝叶斯网络,其假定所有属性是条件独立的。但在实际应用中,条件独立假设往往不成立,因为属性之间常会彼此相关。BN网络可引入隐藏变量,处理数据不完整的问题和变量间复杂的相依性。为此,本文使用了隐朴素贝叶斯网络(Hidden Native Bayesian Network, HNBN)分类器[14]:

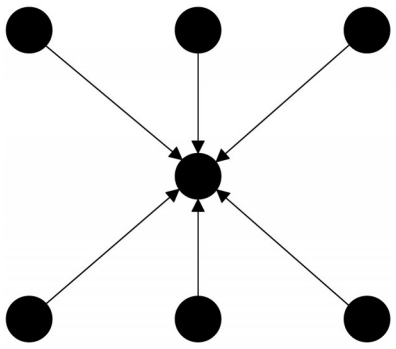
$$c_{\mathrm{HNBN}}=\arg\max_{c\in C}P(c)\prod\nolimits_i P(x_i|x_{\mathrm{hp}_i},c), \tag{13}$$

$$P(x_i|x_{\mathrm{hp}_i},c)=P(c)\sum_{j=1,j\neq i}^{n}w_{ij}P(x_i|x_j,c)。 \tag{14}$$

图2(a)给出了简单的BN网络,图2(b)给出了键入隐藏节点的HNBN网络示意。HNBN模型创建表示每个属性的隐藏父节点层,隐藏父节点(hp_i)代表所有其他属性的加权影响。由此,利用两个属性之间的条件交互关系确定权值w_{ij}。



(a) BN结构



(b) HNBN结构

图2 加入隐藏节点前后的贝叶斯网络对比

Fig. 2 Comparison of Bayesian network before and after adding hidden nodes

1.2.3 决策树

决策树(Decision Tree, DT)算法基于规则集合制定决策。DT学习算法以从上到下递归的方式构建树结构,根节点和中间节点则包含测试条件,利用不同的属性值对实例进行分类,从而为叶节点分配类别标签。DT算法首先从根节点开始,并基于最大信息增益的特征进行数据分割[15]。在迭代过程中,在每个子节点反复进行分割过程,直至每个叶节点的所有实例均属于相同类别。该算法中包含熵(Entropy)和信息增益(IG)的定义。

熵用来衡量系统的不可预测性,计算公式为

$$H(X)=-\sum p(X)\log p(X), \tag{15}$$

式中:$p(X)$为给定类别的部分样本。

信息增益IG指数据集在变换前和变换后的熵

值降低度量，计算公式为

$$G_{\mathrm{I}}(D_{\mathrm{p}},f)=I(D_{\mathrm{p}})-\frac{N_{\mathrm{l}}}{N}I(D_{\mathrm{l}})-\frac{N_{\mathrm{r}}}{N}I(D_{\mathrm{r}}),\quad(16)$$

式中：D_{p}为父节点数据集；f为D_{p}上的部分特征；D_{l}为左侧子节点数据集；D_{r}为右侧子节点数据集；I为噪声；N为总样本数；N_{l}为左侧子节点的样本数；N_{r}为右侧子节点的样本数。

本文使用了极限随机树（Extremely Random Tree，ERT）分类器，通过加入随机性降低了过学习风险[16]。其中，将多个决策树合并，并使用特征的随机子集，基于随机分割的理念来划分节点。令$L=\{(x_1,y_1),\cdots,(x_n,y_n)\}$为训练集，其中样本$\boldsymbol{x}_i=\{f_1,f_2,\cdots,f_n\}$为$D$维向量；$f_i$为特征。ERT创建$M$个独立的决策树。对于每个决策树，$S_p$表示训练集$L$在子节点$p$的子集。然后，对于每个节点$p$，ERT算法基于$S_p$选择最优分支。算法1为ERT分类算法。

算法1 极限随机树算法

输入：训练子集$S_p=\{s_1,s_2,\cdots,sQ_p\}$，样本$s_i=\{f_1,f_2,\cdots,f_n\}$为$D$维向量，$K$为要随机选择的属性数量，$n_{\min}$为节点分割所需的最少样本数。

If $Q_p<n_{\min}$ then

　　停止分割，并将节点定义为叶节点

Else

　　选择随机子分组

End if

for 子分组中的每个特征k do

　　寻找最大值$f_k^{\max}$和最小值$f_k^{\min}$

　　for 子集S_p中的特征k

　　得到$[f_k^{\min},f_k^{\max}]$内的随机切割点f_k^c

　　选择f_k^c作为候选切割

end for

选择分割$f_*<f_*^c$，并使得f_*^c为最小得分

输出：子节点p的最优分割$[f_*<f_*^c]$

1.2.4　人工神经网络

人工神经网络（Artificial Neural Network，ANN）分类器为前向神经网络，一层中的每个节点均连接点下一层中的每个其他节点，图3给出了简单ANN的结构示例[17]。其中，$y=w_ix_i+b$，y为输出值，w_i为权重系数，x_i为输入值，b为偏置系数。

以三层前向ANN为例，中间隐藏层的节点数量计算为$\sqrt{(\alpha+\beta)}+\gamma$，$\alpha$为输入层节点数，$\beta$为输出层节点数，$\gamma$为1和10之间的一个随机均匀数。通常，对输入的加权和应用非线性激活函数，并在训练中通过权值更新将损失最小化。本文在ANN中使用了Sigmoid激活函数。

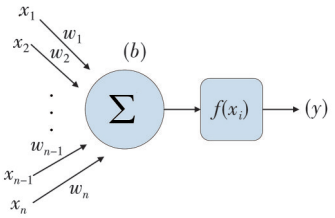


图3　ANN结构示例

Fig. 3　Illustration of the ANN structure

1.3　Ensemble方法

在数据分类预测中，使用基于集成学习理念的Ensemble方法，将不同的弱分类器合并成单个强分类器[18]。通过在每轮训练中增加误分类样本在基础分类器中的权重，实现比单个分类器更好的性能。首先，对所有数据点赋予相同权重。其后，为错误分类的数据点指定更高的权重。由此，高权重数据点将在后续处理中得到重点关注。持续训练模型，直至达到理想的误差值。

令$(x_1,y_1),\cdots,(x_n,y_n)$为已标注训练集，其中，每个$x_i$在实例空间$X$中，每个标签$y_i$在标签集$Y$中。令$D_{l_t}(i)$表示训练实例$I$在第$t$轮的权重。初始时，为所有数据点赋予相同权重，其后在每轮中增加误分类样本在基础分类器中的权重。算法2给出了Ensemble模型的伪代码。本文使用的集成框架属于同质分类器集成，即所有的基础分类器都是相同的，通过不断计算每个基础分类器的误差并对样本和标签分布进行更新来保证多样性，并由此得到集成后的强分类器。

与集成框架中使用不同类型的分类器相比，同质分类器集成具有训练和预测速度快、鲁棒性强、可扩展性强、不易受到数据噪声或异常值影响的优点。异质分类器集成则更擅长处理复杂数据和异质数据。本文实验中使用的在线学习数据集Kalboard360[19]的数据具有较强的规律性，使用同质分类器集成可更好地捕捉模式或规律，提高预测性能。

算法2 集成学习模型

输入：训练数据集$(x_1,\cdots,x_n)$和标签集$(y_1,\cdots,y_n)$

输出：预测结果

为每个样本分配标签$(x_1,y_1),\cdots,(x_n,y_n)$

初始化$D_{I_t}(i)$权重

For $I_t=1,\cdots,T$ do

使用分布$D_{I_t}(i)$训练弱分类器

得到弱分类器预测$h_{I_t}: X\rightarrow\{-1,+1\}$

计算弱分类器误差$\varepsilon_{I_t}=\sum_{h_{I_t}(x)\neq y_i}^{n} D_{I_t}(i)$

更新分布

$D_{I_t}: D_{I_{t+1}}(i)=D_{I_t}(i)\exp(-\alpha_{I_t}y_{I_t}T_{I_t}(x_{I_t}))/C_{I_t}$

$I_t=I_{t+1}$

输出最终预测: $H_{(x)}=sign\left[\sum_{I_t=1}^{T}\alpha_{I_t}h_{I_t}(x)\right]$

算法2中，C_{I_t}为归一化常数，α_{I_t}用于解决过拟合和噪声敏感性问题，即

$$\alpha_{I_t}=\frac{1}{2}\ln\left[\frac{P_{+1}-P_{-1}}{P_{-1}+P_{-1}}\right], \tag{17}$$

式中：P为分类概率估计。

2 实 验

实验将首先评估每个分类器SVM、HNBN、ERT和ANN在单独使用时的预测性能，然后分析应用Ensemble模型后，不同分类器组合的性能变化情况。

2.1 数据集和硬件配置

使用公开在线学习数据集Kalboard360[19]分析所提方法的性能。该数据集中包含480个学习者的不同特征数据，表1给出了用于预测的不同特征的描述。在分类任务中，按照百分制中的0～69，69～79和79～100将学习者分为3类。使用的操作系统为Windows 10，编程语言为Python 3.7，使用sklearn机器学习库。硬件配置为3.2 GHz的Intel i5 4570 CPU，16 GB RAM，NVIDIA GeForce GTX 750 Ti显卡。

表1 数据集特征

Tab. 1 Dataset description

特征	解释	特征类型
籍贯	出生地	类别
性别	学习者性别	类别
生日	出生日期	类别
等级ID	学历水平	类别
年级ID	入学时长	类别
课程	例如数学、计算机等	类别
缺席天数	是否大于7天	类别
访问资源	访问课程内容次数	数值
互动	主动回答问题次数	数值
讨论	参与讨论次数	数值

2.2 评估指标

使用准确度、精度和$F1$指标分析不同分类器的预测性能。准确度为正确分类样本数在总样本数中的占比，计算公式为

$$A=\frac{N_{TN}+N_{TP}}{N_{TN}+N_{TP}+N_{FP}+N_{FN}}, \tag{18}$$

式中：N_{TN}和N_{TP}分别为正确分类的负样本数和正样本数；N_{FP}和N_{FN}分别为错误分类的正样本数和负样本数。

精度用于测量正确分类的正样本数在实际正样本数中的占比，因此能够衡量少数类样本分类的准确度。精度计算公式为

$$P=\frac{N_{TP}}{N_{TP}+N_{FP}}。 \tag{19}$$

召回率为所有预测为正的样本数中正确预测的占比，计算公式为

$$R=\frac{N_{TP}}{N_{TP}+N_{FN}}。 \tag{20}$$

$F1$指标同时考虑到精度和召回率，因此能够更全面地衡量分类器的性能，计算公式为

$$F1=2\times\frac{P\times R}{P+R}。 \tag{21}$$

2.3 实验结果及分析

首先，不使用SMOTE算法和Ensemble框架，仅应用不同机器学习算法对学习者进行分类预测，表2给出了实验结果。从中可发现，SVM算法性能最差，预测准确度仅为68.5%，证明线性分类器不适用于多维数据和不均衡数据集的分类预测。ERT算法取得了92.4%的最高准确度，比次优的ANN算法提升了2.2%的预测准确度，这表明在没有足够训练数据的情况下，神经网络不能有效提取出高质量的分类特征。

表2 单个分类器的性能

Tab. 2 Performance of a single classifier %

分类器	A	P	$F1$
SVM[13]	68.5	66.3	65.4
HNBN[14]	74.9	74.7	76.9
ERT[16]	92.4	91.7	90.8
ANN[17]	90.2	89.8	88.5

图4给出了在应用所提改进SMOTE算法和原SMOTE后，各分类器的准确度性能表现。从表中可发现，原SMOTE算法能够在一定程度上提高数据集的均衡度和分类器的预测性能。其中，ERT分

类器的预测准确度从92.4%提升至93.2%。本文所提SMOTE算法通过生成区分性高的少数类样本，移除边界区分性较弱的样本和噪声样本，在实现数据集平衡的同时，利用高质量的生成样本提高分类器的训练效果，使各分类器均实现了最好性能。ERT分类器的预测准确度比无SMOTE和原SMOTE算法分别提高了1.9%和1.1%。

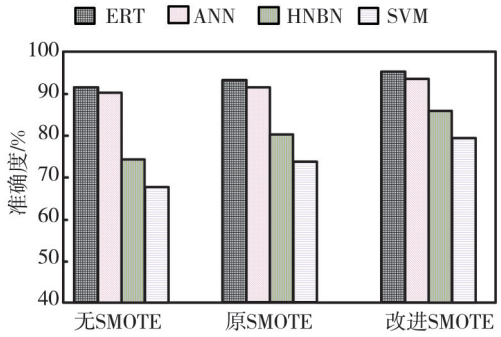


图 4　应用数据集均衡算法后的预测准确度

Fig. 4　Prediction accuracy after applying data equalization algorithms

表 3 给出了应用 Ensemble 模型和改进SMOTE算法后，各分类器的预测性能。从结果中可发现，各模型的性能均实现了显著提升，证明集成学习的理念利用多个弱分类器合并为单个强分类器，并将注意力集中在错分类样本上，能够有效提高最终模型的预测性能。其中，ERT算法依然取得了最好性能，与不使用SMOTE和Ensemble模型的单个ERT分类器相比，预测准确度提高了5.5%。性能提升幅度最大的则是SVM分类器，预测准确度比原单个分类器提升了16.9%。这是因为SVM分类器在集成之前性能较差，在处理数据的不确定性和噪声方面的能力较弱，在应用改进的SMOTE和Ensemble模型后，有效解决了数据噪声和过拟合问题，实现了预测性能的显著改善。

表 3　各分类器结合集成学习后的性能

Tab. 3　Performance of each classifier under ensemble learning framework　%

分类器	*A*	*P*	*F*1
SVM	85.4	84.6	82.9
HNBN	89.8	88.2	87.4
ERT	97.9	97.3	96.8
ANN	95.4	95.5	94.2

此外，图 5 给出了使用改进SMOTE算法和Ensemble模型后，各分类器预测结果的均方误差（Mean Absolute Error，MAE）和均方根误差（Root Mean Squared Error，RMSE）比较。MAE为预测值与真实值之间的平均差，代表预测误差的标准偏差值。RMSE为每个预测值和真实值之间的平方差均值[20]。

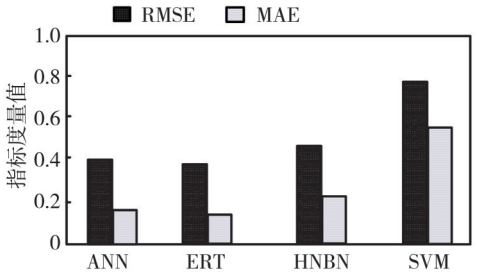


图 5　MAE和RMSE结果

Fig. 5　The results of RAE and RMSE

从图 5 中可发现，ERT算法和ANN算法均能够较好地完成分类和预测任务。其中，ERT的MAE和RMSE值分别为0.14和0.38，表明利用所提SMOTE算法进行数据集均衡化并通过集成学习理念强化分类器性能后，实现了精准的数值预测。

3　结　论

本文提出了基于集成学习理念的机器学习分类预测方法，利用改进SMOTE算法进行数据均衡化处理，并解决了噪声和易混淆样本问题。通过Ensemble算法基于权重调整进行分类器合并，提升了预测性能。在线学习的公开数据集Kalboard360上的实验结果表明，在稀疏多维数据的分类预测任务中，ERT算法在单独使用时得到92.4%的最高准确度。将所提方法应用到ERT算法后，改进的SMOTE算法能够在确保数据集平衡的前提下生成高质量的合成样本，预测准确度比原SMOTE算法提升了1.1%。Ensemble模型通过集成学习的方式有效增强了分类器的性能，实现了97.9%的预测准确度。

参考文献：

[1] 李瑞峰，杨海峰，蔡江辉，等. 一种基于加权深度森林的离群数据挖掘算法[J]. 小型微型计算机系统，2022，43(7)：1426-1431.
LI Ruifeng，YANG Haifeng，CAI Jianghui，et al. Outlier data mining algorithm based on weighted deep forest[J]. Journal of Chinese Computer Systems，2022，43(7)：1426-1431.(in Chinese)

[2] FISCHER C，PARDOS Z A，BAKER R S，et al.

Mining big data in education: affordances and challenges [J]. Review of Research in Education, 2020, 44(1): 130-160.

[3] 刘铁园, 陈威, 常亮, 等. 基于深度学习的知识追踪研究进展[J]. 计算机研究与发展, 2022, 59(1): 81-104.
LIU Tieyuan, CHEN Wei, CHANG Liang, et al. Research advances in the knowledge tracing based on deep learning [J]. Journal of Computer Research and Development, 2022, 59(1): 81-104. (in Chinese)

[4] 陈曦, 梅广, 张金金, 等. 融合知识图谱和协同过滤的学生成绩预测方法[J]. 计算机应用, 2020, 40(2): 595-601.
CHEN Xi, MEI Guang, ZHANG Jinjin, et al. Student grade prediction method based on knowledge graph and collaborative filtering[J]. Journal of Computer Applications, 2020, 40(2): 595-601. (in Chinese)

[5] TRAKUNPHUTTHIRAK R, CHEUNG Y, LEE V C S. A study of educational data mining: Evidence from a thai university[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 734-741.

[6] 申航杰, 琚生根, 孙界平. 基于模糊聚类和支持向量回归的成绩预测[J]. 华东师范大学学报(自然科学版), 2019(5): 66-73.
SHEN Hangjie, JU Shenggen, SUN Jieping. Performance prediction based on fuzzy clustering and support vector regression [J]. Journal of East China Normal University (Natural Science), 2019(5): 66-73. (in Chinese)

[7] 林梦楠, 李金辉. 基于自适应差分进化的学生成绩等级预测神经网络模型[J]. 现代电子技术, 2022, 45(3): 130-134.
LIN Mengnan, LI Jinhui. Student achievement grade prediction model based on neural network optimized by adaptive differential evolution [J]. Modern Electronics Technique, 2022, 45(3): 130-134. (in Chinese)

[8] 张阳, 鲁鸣鸣, 郑一基, 等. 基于图自编码器模型的学生成绩预测[J]. 计算机工程与应用, 2021, 57(13): 251-257.
ZHANG Yang, LU Mingming, ZHENG Yiji, et al. Student grade prediction based on graph auto-encoder model [J]. Computer Engineering and Applications, 2021, 57(13): 251-257. (in Chinese)

[9] DONG X, YU Z, CAO W, et al. A survey on ensemble learning[J]. Frontiers of Computer Science, 2020, 14(1): 241-258.

[10] 李慧芳, 黄姜杭, 徐光浩, 等. 基于多维度特征融合的云工作流任务执行时间预测方法[J]. 自动化学报, 2023, 49(1): 67-78.
LI Huifang, HUANG Jianghang, XU Guanghao, et al. Multi-dimensional feature fusion-based runtime prediction approach for cloud workflow tasks[J]. Acta Automatica Sinica, 2023, 49(1): 67-78. (in Chinese)

[11] KAMAL P, AHUJA S. An ensemble-based model for prediction of academic performance of students in undergrad professional course[J]. Journal of Engineering, Design and Technology, 2019, 17(4): 769-781.

[12] 梅大成, 陈江, 郑涛. 边界与密度适应的SMOTE算法研究[J]. 计算机应用研究, 2022, 39(5): 1478-1482.
MEI Dacheng, CHEN Jiang, ZHENG Tao. Research on SMOTE algorithm based on boundary and density adaptation[J]. Application Research of Computers, 2022, 39(5): 1478-1482. (in Chinese)

[13] SALAWU M D, AROWOLO M O, ABDULSALAM S O, et al. A chi-square-SVM based pedagogical rule extraction method for microarray data analysis [J]. International Journal of Advances in Applied Sciences, 2020, 9(2): 93-100.

[14] CHEN X, YUAN Y, ORGUN M A. Using Bayesian networks with hidden variables for identifying trustworthy users in social networks[J]. Journal of Information Science, 2020, 46(5): 600-615.

[15] YANG F J. An extended idea about decision trees[C]// 2019 International Conference on Computational Science and Computational Intelligence (CSCI), 2019: 349-354.

[16] ACOSTA M R C, AHMED S, GARCIA C E, et al. Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks [J]. IEEE, 2020, 8: 19921-19933.

[17] KURANI A, DOSHI P, VAKHARIA A, et al. A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting [J]. Annals of Data Science, 2023, 10(1): 183-208.

[18] ASHRAF M, ZAMAN M, AHMED M. An intelligent prediction system for educational data mining based on ensemble and filtering approaches[J]. Procedia Computer Science, 2020, 167(1): 1471-1483.

[19] AMRIEH E A, HAMTINI T, ALJARAH I. Mining educational data to predict student's academic performance using ensemble methods[J]. International Journal of Database Theory and Application, 2016, 9(8): 119-136.

[20] HODSON T O. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not[J]. Geoscientific Model Development, 2022, 15(14): 5481-5487.