计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210200110

一种基于 AP-Entropy 选择集成的风控模型和算法

王茂光　杨　行
中央财经大学信息学院　北京 100081
(wangmg@cufe.edu.cn)

摘　要　近年来互联网金融网贷领域涌现出了众多的风控问题,对此采用多种特征选择方法预处理风控领域的数据指标,构建了全面的针对企业信用的风控指标体系,采用 stacking 集成策略研究了基于 AP-Entropy 的信用风险模型。信用风险模型有两层学习器,引入选择集成思想,从种类和数量上筛选基学习器。首先,在 Logistic 回归、反向传播神经网络、AdaBoost 等经典机器学习算法中,采用 AP 聚类算法选出适合企业信用风险的异质学习器作为基学习器;其次,在每次学习器迭代中,利用熵对学习器择优,自动选出 F1 值最高的基学习器,其中改进基于熵的学习器选择算法,提升了基学习器选择过程的效率,降低了模型的计算成本,模型选取 XGBoost 作为次级基学习器。实验结果表明,文中提出的模型和其他模型相比具有更好的学习效果和更强的泛化能力。

关键词:风控指标体系;stacking 集成策略;AP-Entropy 信用风险模型;选择集成;AP 聚类算法;基于熵的学习器选择算法;XGBoost

中图法分类号　TP311

Risk Control Model and Algorithm Based on AP-Entropy Selection Ensemble

WANG Mao-guang and YANG Hang
School of Information,Central University of Finance and Economics,Beijing 100081,China

Abstract　In recent years,many risk control problems have emerged in the field of Internet finance. For this,we adopt a variety of feature selection methods to preprocess data indicators in the field of risk control,and construct a comprehensive risk control indicator system for corporate credit. And we use stacking ensemble strategy to study credit risk model based on AP-entropy. There are two layers of learners in credit risk model. The idea of selection ensemble is introduced to select the base learners from the category and quantity. First,in machine learning algorithms such as Logistic regression,back propagation neural network,AdaBoost,AP clustering algorithm is used to select a heterogeneous learner suitable for corporate credit risk as the base learner. Secondly,in each iteration of the learner,entropy is used to select the best learner,and the base learner with the highest F1 value is automatically selected. Among them,the improved algorithm based on entropy improves the efficiency of base learner selection process and reduces the computational cost of the model. Xgboost is selected as the secondary base learner. The empirical results show that the proposed model has good performance and generalization ability.

Keywords　Risk control feature system,Stacking ensemble strategies,AP-Entropy credit risk model,Selective ensemble,Affinity propagation clustering algorithm,Learner selection algorithm based on Entropy,XGBoost

1　引言

近年小额网贷风险频发,国家逐步减少网贷机构,并成立互联网金融协会加强对网贷机构的监管和规范。其本质上是要加强对网贷企业风控的研究,建立健全网贷企业信用风险体系。网贷企业信用风险主要包含企业的信贷违约风险。

目前比较成熟的企业信用风险模型是 SVM 等单一机器学习模型。但单一学习器受限于实际情况,训练结果往往不稳定,呈现弱监督性。集成学习就是组合弱监督模型得到强监督模型,因此集成学习在网贷企业信用评估领域的应用效果优于单一机器学习模型[1]。同时,集成学习按照学习器种类的不同,可分为同质学习器集成和异质学习器集成。同质学习器的集成学习模型已经有较多应用[2-4],异质学习器的应用以及集成学习器种类的选择鲜有研究。

本文针对网贷企业风控问题,提出一种基于 AP 聚类(Affinity Propagation Clustering Algorithm)和熵的多学习器选择集成模型,简称 AP-Entropy 选择集成模型。该模型的构建包括两部分:第一部分,确定基学习器种类和个数,将信用风险领域典型的单一学习器输入模型,利用 AP 聚类算法选出效果显著且彼此存在差异的异质基学习器,其次穷举得到最优学习器个数,利用熵对基学习器进行二次筛选;第二部分,选用适合网贷企业信用风险的次级学习器,采用 Stacking 集成策略对模型进行集成。本文第 2 节主要描述当前企业信用风险领域的发展现状及常用模型;第 3 节主要讲述 AP-Entropy 模型核心算法;第 4 节重点描述模型搭建及对实验结果的分析;最后总结全文并展望未来。

通信作者:杨行(yanghangv@163.com)

2 相关工作

2.1 单一学习器

有效的企业信用风险评估能够识别信用风险并将其降低至可控范围，对企业自身和投资者均能起到预警作用。常用的企业信用风险评估方法有要素分析法、财务指标分析法、传统统计学方法、机器学习方法等。随着数据量的爆炸式增长以及计算速度和存储空间的提升，机器学习算法成为各行各业研究和使用的热点。企业信用风险评估领域也不例外，典型的有 Logistic、SVM、神经网络、决策树及相关简单集成模型。单一模型中，Logistic 模型由于简单稳定及模型可解释性较好已经被广泛用在信用风险领域，许多学者还对模型做出改进。如 Fang 等人从变量间的网格关系入手，改进传统 Logistic 模型，通过惩罚方法同时实现变量选择和参数估计，实验证明改进后的模型更具有应用价值[5]。Zhang 等[6]基于 logistic 和 svm 混合预警模型研究商业银行中客户违约的线性和非线性的复杂特征。Li 等[7]采用反向传播神经网络模型拟合网络信用环境下对网贷借款人的信用风险。

决策树包含分类树和回归树，可分别对离散变量和连续变量做预测回归，因此非常适用于金融领域数据。Liu 在构建企业信用风险指标体系后，选择经典决策树 C4.5 算法构建评估模型，并取得了较好的实验结果[8]。决策树相关的应用在信用评估领域已经比较成熟，但由于单棵树构建的模型往往不够稳定，又或者在训练过程中容易出现过拟合现象，许多学者还引入随机森林和一些以决策树为基学习器的简单集成模型。Yu 借助随机森林研究 P2P 网贷企业中样本不均衡问题[9]。极端随机树(Extremely Randomized Trees，ET)与随机森林均由许多决策树构成，随机森林采用类 bagging 策略选择样本进行训练，而 ET 树可以使用所有样本进行训练，往往在信用风险领域可以取得更好的结果[10]。梯度提升树 Gradient Boosting Machine(GBM)算法是基于梯度下降算法得到的提升树模型，GBM 算法经常作为独立模型应用到信用风险评估领域，并取得了较好的效果。GBM[11]使用基于 histogram 的决策树算法，在内存和计算上进行了改进。Fei 等人借助 GBM 算法，采用 Stacking 策略集成模型，实验结果表明 GBM 算法的 AUC 和准确率较高[12]。AdaBoost 常常选择决策树作为弱学习器，充分考虑每个分类器的权重，提升了分类和预测效果，近年来也被广泛应用于信贷问题中[13]。

2.2 多学习器集成

集成学习作为机器学习的分支之一，近些年异常火热，同时集成学习在企业信用风险应用方面也取得了较大的进展。Yu 等人利用 Bagging 模型增加样本空间的多样性生成不同的数据集，利用极限学习机作为基学习器在不同的数据集上训练，其次利用 Stacking 策略，将有较高学习能力的 DBN(深度信念网络)作为次级学习器，对每个极限学习机生成的结果进行融合[14]。在增强样本多样性的方面，Chen 等人采用 Adaboost 方式，通过对样本赋予不同的权重，提升了多样性。实证研究表明，基于 SVM 的随机子空间和 Adaboost 集成方法是一种企业信用风险评估的有效模型[15]。在基学习器选择方面，大部分研究多基于同质学习器，仅有少数的研究关注异质学习器的集成。基于异质学习器集成的优势在于不同类型分类器对相同数据可以学得不同的观点并可以相互补充。Nascimento 等[16]采用十多种不同的学习器构建集成模型，增加了基学习器输出的多样性。Ala'raj 等[17]提出了一种新的分类器-共识系统对集成模型中的异质学习器进行融合，与传统的 LR\MARS 模型结合策略相比，该方法模拟了真实的信用评估专家团队的行为。Xia[18]通过 Bagging 和 Stacking 策略构建了一种新奇的异质集成信用风险评估模型。该模型在模型生成、基学习器选择和结合策略上不同于以往的集成模型，实验结果证实了该模型的优越性。由此可见，选择异质学习器较同质学习器而言有较大的优势。

另一方面，对集成学习而言，基学习器的选择也至关重要，随着基学习器数量的增多，伴随出现准确率提升不明显甚至下降、计算效率低下的问题。针对这些问题，Zhou 等提出选择集成学习[19]，并根据这些年的研究将选择集成学习可分为以下几类[20]：基于聚类的选择性集成，基于排序的选择性集成，基于选择的选择性集成。Chen 等人提出利用变相似度聚类技术和贪婪算法来进行选择[21]，Zheng 等[22]对训练后的基分类器进行聚类选择，将相似基分类器剔除，不仅减少了内存占用空间，且可以提高效率。Chen 等[23]提出一种多层次选择性集成学习算法，即在基分类器中通过多次按权重进行部分选择，形成多个集成分类器，对形成的集成分类器进行再集成，最后通过对多个集成分类器多数投票的方式决定算法的输出。综合来看，选择性集成学习目前在筛选基学习器的个数、提高模型准确率及减少算法计算量这一领域已经比较成熟，但很少有统一的准则确定学习器的种类。

3 基于 AP-Entropy 选择集成模型

集成学习常通过集成大量学习器来获取比单一模型更好的性能。然而这种做法一方面需要大量的计算和存储空间，容易导致运算速度变慢；另一方面，当个体学习器的数目变多后，学习器之间的差异性减小，学习效果下降。因此，如何筛选学习器个数并选出异质学习器就成为了研究的重点，这也是 AP-entropy 选择集成模型所要解决的问题。首先，因为本文将集成信用风险领域的多种不同模型，这些模型之间是异质的，异质学习器集成采用 Stacking 集成策略得到的效果最好，bagging 和 boosting 集成框架多集成同质弱学习器，因此本文选取 Stacking 集成策略，整体采用两层模型堆叠。第一层是多个单一学习器组成的集合，第二层决定采用一个强学习器。在第一层学习器集成的过程中，本文先利用 AP 聚类算法从种类上筛选学习器，初步选出异质性较高的个体学习器，保证基学习器之间的差异性，其次对这些类型的学习器进行穷举，选出最佳的学习器个数，再用基于熵的学习器选择算法对基学习器进行二次筛选，进一步选出学习器中差异性大的学习器，同时减少学习器数量，节省计算机内存开销，在保证学习效果的情况下提升模型运行效率和泛化能力。对于次级学习器，本文将通过实验确定。模型整体框架如图 1 所示。

AP-Entropy 选择集成模型的核心是基于 AP 聚类的学习器选择算法和基于熵的学习器选择算法。前者主要用于选出基学习器的种类，后者用于筛选基学习器的数量。

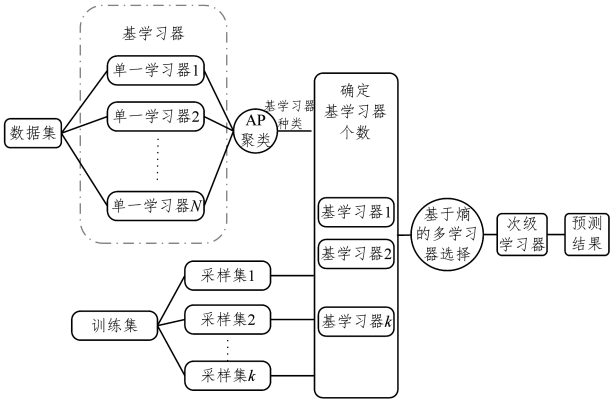

图 1 AP-Entropy 选择集成模型

Fig. 1 AP-Entropy selective ensemble model

3.1 基于 AP 聚类的学习器选择算法

无监督聚类算法是选择集成学习中选择分类器的策略之一。聚类算法将基学习器分成几个簇，在每个簇中选出分类效果最好的分类器用于集成，通过簇保证不同簇分类器的差异性，同时剔除同一簇内其他的分类器，减少学习器个数。这种方法需要注意的是初始基学习器个数的选择和簇内距离定义。

为了保证模型的泛化能力，经过多年的实验研究，选取决策树、Logistic、SVM、反向传播网络、随机森林、ET 树（极端随机树）、AdaBoost、GBM（梯度提升树）这 8 种信用风险领域典型的单一学习器。为了解决簇内距离定义和簇内最优分类器选择的问题，本文引入 AP 聚类算法。AP 聚类算法假设所有数据点作为簇中心，采用亲和度传播算法（Affinity Propagation）得到聚类中心点。通过 AP 聚类算法选出的聚类中心点是数据中实际存在的点，同时也是该簇内最优的点，因此该点对应的分类器为本簇内的最优分类器。其中，亲和度传播算法可描述为[24]：

（1）将吸引度矩阵 $\boldsymbol{R}$ 和归属度矩阵 $\boldsymbol{A}$ 初始化为全 0 矩阵。

（2）更新吸引度矩阵 $\boldsymbol{R}$。

$$r_{t+1}(i,k)=\begin{cases}S(i,k)-\max\{a_t(i,j)+r_t(i,j)\}, & i\neq k\\ S(i,k)-\max\{S(i,j)\}, & i\neq k\end{cases} \tag{1}$$

（3）更新归属度矩阵 $\boldsymbol{A}$。

$$a_{t+1}(i,k)=\begin{cases}\min\{0,r_{t+1}(k,k)+\sum\limits_{j\neq i,k}\max\{r_{t+1}(j,k),0\}\}, & i\neq k\\ \sum\limits_{i\neq k}\max\{r_{t+1}(j,k),0\}, & i=k\end{cases} \tag{2}$$

（4）根据衰减系数 λ 对两个公式进行衰减。

$$r_{t+1}(i,k)=\lambda * r_t(i,k)+(1-\lambda) * r_{t+1}(i,k) \tag{3}$$

$$a_{t+1}(i,k)=\lambda * a_t(i,k)+(1-\lambda) * a_{t+1}(i,k) \tag{4}$$

其中，吸引度（responsibility）矩阵 $\boldsymbol{R}$ 中 $r(i,k)$ 描述了数据对象 k 适合作为数据对象 i 的聚类中心的程度，表示的是从 i 到 k 的消息；归属度（availability）矩阵 $\boldsymbol{A}$ 中 $a(i,k)$ 描述了数据对象 i 选择数据对象 k 作为其据聚类中心的适合程度，表示从 k 到 i 的消息。AP 聚类算法的本质就是选取吸引度和归属度之和（$a+r$）最大的值作为聚类中心。

本文采用 AP 聚类算法对学习器种类进行选择，算法具体步骤如下：

Step1 将数据集输入至 8 个单一学习器组成的基学习器集合中，得到每个单一学习器的学习指标，指标包含准确率、F1 值等。

Step2 准确率作为评判单一学习器好坏的第一级标准，筛选得到准确率 85%以上的单一学习器，并将筛选出的单一学习器预测结果及对应单一学习器组合成矩阵 $\boldsymbol{D}$。

Step3 将 $\boldsymbol{D}$ 输入 AP 聚类算法，得到 k 个聚类中心。因为 AP 聚类算法预测得到的聚类中心是实际存在的点，这里的得到的聚类中心代表单一学习器，因此筛选得到 k 个单一学习器。将这 k 个单一学习器作为新的基学习器。

AP 聚类算法的最终结果见表 1。

表 1 AP 聚类结果

Table 1 AP clustering results

索引	模型	标签	是否为簇中心
0	决策树	0	是
1	Logistic	1	是
2	SVM	2	是
3	BP 神经网络	3	是
4	随机森林	4	否
5	极端随机树	4	是
6	AdaBoost	5	是
7	梯度提升树	6	是

从表 1 可以看出，进行 AP 聚类后，8 个学习器被分成 7 簇，选出每个簇的中心作为本簇内的最优分类器，在标签为 4 的分类器中，根据 $a+r$ 值的大小，保留极端随机树模型，删除随机森林模型，简化基学习器的种类。同时，我们分别采用随机森林和极端随机树对数据进行单独训练，结果发现随机森林出现过拟合现象，且 F1 值、召回率等指标低于极端随机树，这进一步证明了 AP 聚类结果的有效性。

3.2 基于熵的学习器选择算法

基于熵的学习器选择是基于“熵”来衡量基学习器集合的差异性，“熵”反映不确定程度，当集合中的基学习器都将某个样本正确划分或者错误划分时，集合的差异性最小，熵值为 0。熵值的范围为[0,1]，熵越大，代表基学习器集合的差异性越大。非成对熵度量公式[25]如下所示：

$$E=\frac{1}{N}\frac{1}{L-\lfloor L/2\rfloor-1}\sum_{i=1}^{N}\min\{l(x_i),(L-l(x_i))\} \tag{5}$$

其中，L 代表分类器的个数，N 代表样本个数，$l(x_i)$ 代表将样本 x_i 分类正确的分类器的个数。

3.1 节中确定了学习器的种类为 7 种，本文对这 7 种学习器进行 n 次迭代，即每次迭代中选取 F1 值最大的学习器，最后得到 n 个学习器，对这 n 个学习器采用熵进行筛选，n 的确定过程见 4.2.1 节。本文利用前向逐步搜索法来最大化集合的差异性，从空集合开始，在未加入基学习器集合中挑选一个基学习器，并计算目标函数，即整个集合的熵值。若该基学习器的加入能够使目标函数增加，则放入该基学习器；若目标函数不能增加，则不放入该学习器，直至对所有的基学习器都已经做出判断。因此我们在式（5）的基础上改进熵度量公式[26]：

$$E=\frac{1}{N}\frac{1}{L-\lfloor L/2\rfloor}\sum_{i=1}^{N}\min\{l(x_i),(L-l(x_i))\} \tag{6}$$

因为本文利用前向逐步搜索法来度量集合差异性，我们

设计的基分类器集合中所包含的基学习器的个数至少为 2。基于熵的学习器选择算法的具体步骤如下：

Step1 从根据 F1 值高于 60%的原则筛选出的基学习器集合 M 中随机挑选一个基学习器作为初始成员，将其放入选择性集成学习集合 S 中，并计算整个集合的熵值 $E1$；

Step2 从 M 中未被挑选的个体学习器中随机挑选一个基学习器加入选择性集成学习集合 S 中，并计算此时整个集合的熵值 $E2$，若 $E2 > E1$，则保留此基学习器，否则，从 S 中删除此学习器；

Step3 重复 Step2，直至 M 中所有的个体学习器均被遍历；

Step4 生成的选择性集成学习集合 S 作为后续集成学习的基学习器集合。

本文选取迭代次数 $n=50$（第 4 节将对 n 的选取做出说明），即有 50 个基学习器。经 Step1－Step4 训练后，模型中有 13 个 ET 树、12 个 GBM 模型、8 个 Logistic 模型、8 个 AdaBoost 模型、7 个决策树模型和 1 个反向传播神经网络模型。经过熵的筛选，SVM 学习器在增加后熵值明显减小，不符合算法标准，被剔除。

4 实验结果及分析

4.1 特征选择

本文数据集来源于 2019 年网贷之家平台上相关的小额网贷企业相关信息，共包含 27 个特征，5570 条数据。本文将数据集划分成训练集、测试集和验证集，比例为 2∶1∶1。表 2 列出本文数据特征及标号。鉴于仅有 9 条数据缺失，本文对缺失数据直接进行删除。通常情况下，特征中都会包含预测能力较弱的特征和相关性较高的特征，针对这些特征需要进行特征选择。

表 2 数据特征

Table 2 Data characteristics

特征	标号	特征	标号
参考收益	x_1	保障模式	x_{14}
投资期限/月	x_2	资金净流入/万元	x_{15}
综合评分	x_3	平均借款期限	x_{16}
点评人数	x_4	投资人数/个	x_{17}
注册年限	x_5	人均投资金额/万元	x_{18}
运营时间	x_6	待收投资人数/人	x_{19}
待还余额/万	x_7	借款人数/人	x_{20}
昨日成交量	x_8	人均借款金额/万元	x_{21}
关注投友数	x_9	借款标数/个	x_{22}
注册资金/万元	x_{10}	待还借款人数/人	x_{23}
银行存管	x_{11}	ICP 认证	x_{24}
自动投标	x_{12}	平台背景	x_{25}
债券转让	x_{13}	加入协会	x_{26}

4.1.1 基于 XGBoost 的特征选择

XGBoost 算法本质上是不断添加树，来拟合上次的残差，当训练完成得到 k 棵树时，将每棵树对应的分数加起来就是该样本的预测值。因此，XGBoost 可以用于特征筛选，具体方法是通过被筛选特征在每棵树中的分裂次数的和去计算每个特征的得分[19]。本文选用 XGBoost 算法进行特征选择，用特征得分来表示这个特征的重要性，每个特征的得分见图 2，得分越高表明该特征对数据集的贡献越大。

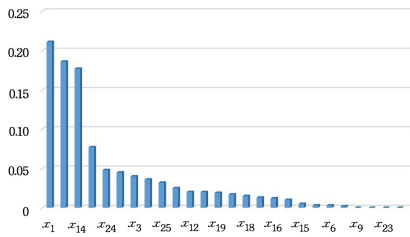

图 2 基于 XGBoost 的特征选择

Fig. 2 Feature selection based on XGBoost

根据图 2 按特征重要性得分排序后的特征，由于注册年限与运营时间线性相关，运营时间的特征重要性略高于注册年限，因此选择运营时间，而筛选掉注册年限，最后保留特征重要性在 0.003 以上的指标，共 21 个。

4.1.2 基于信息增益的特征选择

信息增益（Kullback-Leibler Divergence）常常用来衡量一个特征对整体的贡献程度，经常用于特征选择。信息增益的基础是熵，一种度量随机变量不确定性的指标。熵可以细化为信息熵和条件熵，计算公式如下[27]：

$$IG(T)=H(S)-H(C|T) \tag{7}$$

通过编写 python 程序，可以得到整体数据集的熵和每个特征值的信息增益，并且信息增益值越大，表明该特征对整体数据集的贡献越大。经计算，每个特征的信息增益见图 3。

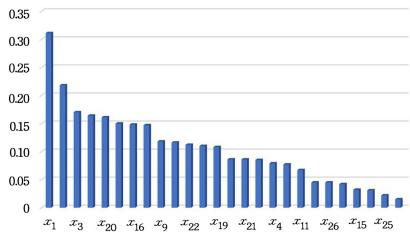

图 3 每个特征的信息增益

Fig. 3 Information gain of each feature

根据图 3 按信息增益排序后的特征，删除信息增益值小于 0.04 的特征，保留 22 个特征。

4.1.3 基于 WOE/IV 的特征选择

特征选取方法中，IV（Information Value）值衡量特征的预测能力，IV 值的计算以 WOE（Weight of Evidence）值为基础。WOE 值越大，该特征的预测效果就越好，但是同时对于每个样本的每个变量来说 WOE 值包含正负，如果要衡量整个特征的预测能力，采用 WOE 值可能出现正负抵消的情况，使得整体预测能力大打折扣。为了弥补 WOE 的不足，本文选取基于 WOE 计算的 IV 值，IV 值越大，该特征的预测能力越强。图 4 是计算后的每个特征的 IV 值按大小排序后的结果。

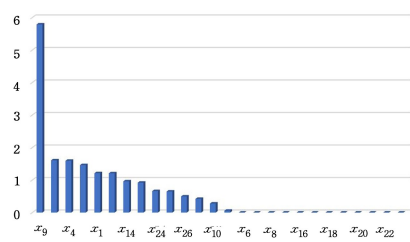

图 4 每个特征的 IV 值

Fig. 4 Information value of each feature

根据 IV 值的筛选标准，本文删除没有预测能力的特征（$IV<0.02$）和预测能力可疑的特征“关注投友数”，保留 13 个特征。从图 4 中可以看出，有一些特征的 IV 值过小，这可能是由于每个样本 WOE 值过小导致出现了极端情况。

4.1.4 特征选择对比

本文实验采取 3 种特征选取方式，对这 3 种特征选取方式进行分别考虑和两两组合考虑，两两组合中删除两种特征选取方式共同剔除的特征，整理得出 6 个特征空间，分别是基于 XGBoost 选取的特征空间 V1、基于信息增益选取的特征空间 V2、基于 WOE/IV 选取的特征空间 V3、基于 WOE/IV & XGBoost 选取的特征空间 V4、基于 WOE/IV & 信息增益选取的特征空间 V5、基于信息增益 &XGBoost 选取的特征空间 V6，其中由于基于信息增益的特征选取方式和基于 XGBoost 的特征选取方式没有共同剔除的特征，故 V6 样本空间等于原样本空间，没有产生新的特征空间，本文不再考虑。分别将这 5 个特征空间输入到本文构建的模型中，各项指标结果见表 3。图 5 给出特征空间的 10 个公共特征，这些指标对于网贷企业信用风险评估具有重要的参考意义。

表 3 不同特征空间的结果输出

Table 3 Result output of different feature spaces

特征空间	*F1*	*Accuracy*	*Auc*	*Precision*	*Recall*
V1	0.93	0.91	0.94	0.92	0.95
V2	0.94	0.92	0.97	0.97	0.92
V3	0.90	0.87	0.91	0.89	0.90
V4	0.93	0.91	0.94	0.92	0.95
V5	0.93	0.91	0.94	0.92	0.95

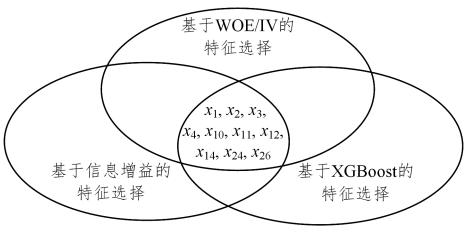

图 5 公共特征

Fig. 5 Common features

从表 3 中结果可以看出，将 V2 输入模型后，预测效果较好，因此本文选取基于信息增益的特征选取方法。V2 = {‘ICP’,‘人均借款金额/万元’,‘人均投资金额/万元’,‘保障模式’,‘借款人数/人’,‘加入协会’,‘参考收益’,‘平均借款期限/月’,‘待收投资人数/人’,‘待还余额/万元’,‘投资期限/月’,‘昨日成交量’,‘注册资金/万元’,‘点评人数’,‘自动投标’,‘银行存管’,‘关注投友数’,‘注册年限’,‘借款标数(个)’,‘待还借款人数/人’,‘投资人数/人’}。其中 V3 由于极端情况的出现导致删除过多特征，输入模型后的预测效果稍弱。这也从侧面反映，在进行特征选取时，要注意选取多种方法进行比较，避免极端情况的出现影响后续模型效果。

4.2 信用风险模型构建

本文使用 Stacking 集成策略构建两层集成模型。

在基学习器的选择上，本文首先采用 AP 聚类算法，从典型的信用风险评估模型中选出 7 个差异性较大的模型（决策树、Logistic、SVM、反向传播神经网络、极端随机树、AdaBoost、梯度提升树），确定本文基学习器的种类；其次，通过实验选择合适的迭代次数（具体见 4.2.1），迭代基学习器，采用熵对迭代后的基学习器进行选择，最终选出 49 个基学习器。这样不仅进一步增大了学习器之间的差异性，而且减少了学习器的个数，降低了计算机内存开销。

在次级学习器的选择上，Logistic 回归因为简单和易于实现常作为次级学习器，但是由于基学习器获得的结果与因变量具有高度相关性，应用到 Logistic 回归模型时容易出现多重共线性问题。而 XGBoost 这种基于树的算法不受共线性问题的影响，适合作为次级学习器[28]。因此，本文选用 XGBoost 作为次级学习器。

4.2.1 实验

在使用 AP 算法对初级学习器进行种类上的选择后，需要对学习器迭代 n 次，即学习器数量为 n 个。我们经过大量实验，发现在多学习器集成领域，$n=50$ 为学习器数量的基线，因此本文从[50,300]之间对学习器个数进行穷举，本文罗列几个具有代表性的 n，其具体的学习指标见表 4。虽然各项衡量指标小幅度变化，但是相对来说基学习器个数取值为 50 时，既保证了准确率又可以降低计算开销和存储空间。因此，设定原始学习器集合中基学习器个数为 50。为了方便对比，运行耗时为在实验室同一台机器下测试的结果。

表 4 不同基学习器个数下的模型指标

Table 4 Model index under different number of base learners

基学习器个数 n	大约耗时/min	*F1*	*Accuracy*	*Auc*	*Precision*	*Recall*	*Error*1	*Error*2
$n=50$	3	0.938	0.920	0.971	0.961	0.915	0.070	0.084
$n=100$	15	0.939	0.922	0.974	0.958	0.920	0.076	0.079
$n=150$	39	0.938	0.920	0.974	0.957	0.920	0.078	0.079
$n=200$	71	0.938	0.920	0.959	0.959	0.917	0.074	0.082
$n=250$	126	0.939	0.922	0.974	0.961	0.918	0.071	0.082
$n=300$	253	0.938	0.920	0.974	0.959	0.918	0.075	0.082

4.2.2 模型对比

为证明本文 AP-entropy 选择集成模型的泛化能力，将 AP-entropy 选择集成模型与典型单一模型（Logistic、BP 神经网络）、基于 AP 的选择集成模型和基于熵度量的选择集成模型进行对比，比对的标准是模型的 F1 值、准确率、召回率、精度、Auc 值。表 5 即为模型在测试集下的性能比较。

表 5 不同模型在测试集下的性能比较

Table 5 Performance comparison of different models under test set

模型	*F1*	*Accuracy*	*Auc*	*Precision*	*Recall*
反向传播神经网络	0.868	0.844	0.935	0.978	0.780
Logistic	0.875	0.819	0.934	0.801	0.964
AP-选择性集成学习	0.938	0.920	0.971	0.961	0.915
基于熵度量的选择性集成	0.934	0.916	0.972	0.969	0.902
AP-entropy 选择集成模型	0.940	0.924	0.973	0.968	0.916

通过表 5 可以看出，集成后的模型比单一模型的性能更好，且各项性能指标更加稳定。本文提出的 AP-Entropy 模型比使用单一选择集成方法性能更好。

为了进一步证明本文模型的泛化能力，在 prosper 在线 P2P 借贷平台数据集[29]和 Leading Club 数据集[30]上应用本模型，并将本文模型与典型机器学习模型进行对比，部分实验结果见表 6、表 7。从表中可以看出，本文提出的多学习器选

择集成模型较单一机器学习模型在各项学习指标上均有提升，各项指标的表现也更加稳定，因此本文提出的模型对企业信用风险的研究具有参考意义和应用价值。

表 6 prosper 数据集下不同模型的性能比较

Table 6 Performance comparison of different models in prosper dataset

模型	*F*1	*Accuracy*	*Auc*	*Precision*
ap_entropy_xgboost	0.847	0.740	0.701	0.744
朴素贝叶斯	0.838	0.729	0.576	0.744
神经网络	0.846	0.733	0.584	0.733
Logistic	0.845	0.732	0.634	0.733
SVM	0.846	0.733	0.436	0.733

表 7 Leading club 数据集下不同模型的性能比较

Table 7 Performance comparison of different models in Leading club dataset

模型	*F*1	*Accuracy*	*Auc*	*Precision*
ap_entropy_xgboost	0.974	0.959	0.949	0.952
朴素贝叶斯	0.957	0.934	0.898	0.949
神经网络	0.884	0.80	0.701	0.800
Logistic	0.961	0.938	0.892	0.934
SVM	0.952	0.923	0.872	0.915

4.2.3 假设检验

为了让实证结果更加可信，本文利用 McNemar 检验对不同模型进行显著性检测，结果如表 8 所列。其中 * * 代表显著性水平在 0.01 下模型之间具有差异性，* 代表显著性水平在 0.1 下模型之间具有差异性。

表 8 不同模型的 McNemar 检验结果

Table 8 McNemar test results of different models

模型	反向传播神经网络	Logistic	AP-选择性集成学习	基于熵度量的选择性集成
AP-entropy 选择集成模型	166 * *	141 * *	3.17 *	5 *
基于熵度量的选择性集成	5 *	153 * *	3 *	
AP-选择性集成学习	184 * *	154 * *		
Logistic	189 * *			

从表 8 中可以看到，集成模型和单一机器学习模型有显著区别，且 AP-entropy 选择集成模型与 AP-选择性集成学习及基于熵度量的选择性集成均有明显区别，模型存在差异。综合表 5，本文提出的 AP-entropy 选择集成模型在 5 个评价指标性能上均有提升。对于大规模的借贷业务而言，评分模型的精确度即使提高 1%，也会给银行业和金融机构带来显著的收益。

结束语 本文提出 AP-Entropy 选择性集成模型，利用不同的特征选择方法，得到全面的企业信用风险指标体系。在基学习器选择中创新性引入 AP 聚类算法和熵对学习器的种类和个数进行筛选，从而达到减少基学习器数量，增大基学习器的异质性这一目标，并降低计算和内存开销。模型采用 stcking 集成策略构建两层模型。实验各项指标表明，本文模型有良好的学习效果和泛化能力，能给网贷企业和投资者提供一定的预警价值。

本文的模型也有一些不足之处。在对基学习器的选择中，不论是种类还是数量上的选择，本文都是采取静态方式完成，未来将尝试使用动态选择方式。同时，未来还将在模型中加入规则策略，扩大模型的适用场景，并尝试从模型可解释性入手对模型进行解释。

参 考 文 献

[1] YAN R J, YIN S Q. Micro-blog credit evaluation model based on selective neutral network ensembl[J]. Computer Engineering and Design, 2018, 377(5): 286-291.

[2] YANG J, YUAN Y L, YU H L. Selective Ensemble Learning Algorithm of Extreme Learning Machine Based on Ant Colony Optimization[J]. Computer Science, 2016(43): 266-271.

[3] LIU J P, HE J Z, MA T Y. Selective Ensemble of KELM-Based Complex Network Intrusion Detection[J]. Acta Electronica, 2019, 47(5): 1070-1078.

[4] HU X J, KANG N. SVM selective ensemble learning method based on feature selection[J]. Electronic Technology & Software Engineering, 2019(18): 143-144.

[5] FANG K N, FAN X Y, MA S G. Forecasting of Enterprise's Credit Risk Based on Network-logistic Model[J]. Statiscal Research, 2016, 33(4): 50-55.

[6] ZHANG Q, HU L Y, WANG Y. Study on credit risk early warning based on Logit and SVM[J]. System Engineering-Thery & Practice, 2015(7): 1784-1790.

[7] LI X, DAI Y C. Research on Early Warning Model of Banking Credit Risk Based on Logit and SVM[J]. Wuhan Finance Monthly, 2018(2): 33-37.

[8] LIU Y. The Application of Decision tree algorithm in credit risk assessment of P2P new loan[D]Changsha: Hunan University, 2016.

[9] YU X H, LOU W G. P2P Online Loan Credit Risk Evaluation, Early Warning and Empirical Research Based on Random Forest [J]. Financial Theory & Practice, 2016, 439(2): 53-58.

[10] PIERRE G, ERNST D, WEHENKEL L. Extremely randomized trees[J]. Machine Learning, 2006, 63(1): 3-42.

[11] ALEXEY N, ALOIS K. Gradient boosting machines, a tutorial [J]. Frontiers in Neurorobotics, 2013, 7: 21.

[12] FEI H Y, HUANG H. Research on Internet Credit Risk Prediction Based on Model Fusion[J]. Statistics and Applications, 2019, 8(5): 12.

[13] ZHOU Q Y. Application Research of Improved AdaBoost Algorithm in Credit Imbalance Classification[D]. Hangzhou: Zhejiang Gongshang University, 2020.

[14] YU L, YANG Z, TANG L. A novel multistage deep belief network based extreme learning machine. ensemble learning paradigm for credit risk assessment[J]. Flexible Services & Manufacturing Journal, 2016, 28(4): 576-592.

[15] CHEN Y, SHI S, PAN Y, et al. Hybrid ensemble approach for credit risk assessment based on SVM[J]. Computer Engineering and Applications, 2016(4): 115-120.

[16] NASCIMENTO D, COELHO A, CANUTO A. Integrating complementary techniques for promoting diversity in classifier ensembles: A systematic study[J]. Neurocomputing, 2014, 138: 347-357.

[17] ALA'RAJ M, ABBOD M F. Classifiers consensus system approach for credit scoring[J]. Knowledge-Based Systems, 2016, 104: 89-105.

(下转第 80 页)

Parallel and Distributed Computing, 2019, 133(11): 93-106.
[2] UTAMIMA A, REINERS T, ANSARIPOOR A H, et al. Optimisation of agricultural routing planning in field logistics with evolutionary hybrid neighbourhood search[J]. Biosystems Engineering, 2019, 184(8): 166-180.
[3] YU H Z, LU F. A multi-modal multi-criteria route planning method based on genetic algorithm[J]. Acta Geodaetica et Cartographica Sinica, 2014, 43(1): 89-96.
[4] GUO C Z, MEGURO J, KOJIMA Y, et al. Automatic lane-level map generation for advanced driver assistance systems using low-cost sensors[C]// 2014 IEEE International Conference on Robotics and Automation. Hong Kong, 2014: 3975-3982.
[5] JIANG K, YANG D G, LIU C R, et al. A flexible multi-layer map model designed for lane-level route planning in autonomous vehicles[J]. Engineering, 2019, 5(2): 305-318.
[6] DENG Y, CHEN Y X, ZHANG Y J, et al. Fuzzy Dijkstra algorithm for shortest path problem under uncertain environment [J]. Applied Soft Computing, 2012, 12(3): 1231-1237.
[7] ZHENG N B, LU G, LI Q Q, et al. The adaption of A* algorithm for least-time paths in time-dependent transportation networks with turn delays[J]. Acta Geodaetica et Cartographica Sinica, 2019, 5(2): 93-100.
[8] QIN F, WU J, ZHANG X F, et al. Improved Search Algorithm Based on A* for Bidirectional Preprocessing[J]. Computer Systems & Applications, 2019, 28(5): 95-101.
[9] XIAO P, ZHOU Z F, ZHAO Y. Discussion on uninterrupted navigation of agricultural machinery based on SINS/GNSS[J]. Journal of Navigation and Positioning, 2019, 7(1): 33-37.
[10] CHEN H Y, ZHANG Y. An Overview of Research on Military Unmanned Ground Vehicles[J]. Acta Armamentarii, 2014, 35(10): 1696-1706.
[11] LI Z F, YANG Y J, WANG X. Rule based shortest path query algorithm[J]. Journal of Software, 2019, 30(3): 515-536.
[12] GUO X Y, LUO X. Global Path Search based on A* Algorithm [C]// International Conference on Transportation & Logistics, Information & Communication, Smart City (TLICSC 2018). Chengdu, China, 2018: 369-374.
[13] REN T Z, ZHOU R, XIA J, et al. Three-dimensional path planning of UAV based on an improved A* algorithm[C]// 2016 IEEE Chinese Guidance, Navigation and Control Conference. Nanjing, China, 2016: 140-145.
[14] ARTIGUES C, HUGUET M J, GUEYE F, et al. State-based accelerations and bidirectional search for bi-objective multi-modal shortest paths[J]. Transportation Research Part C: Emerging Technologies, 2013, 27(1): 233-259.

CAO Bo, born in 1994, master. His main research interests include construction and application of high precision map.

LI Yong-le, born in 1984, Ph.D supervisor. His main research interests include omnidirectional vision, virtual reality and computer vision.

(上接第 76 页)

[18] XIA Y F. A novel heterogeneous ensemble credit scoring model based on bstacking. approach[J]. Expert Systems with Applications, 2018, 93.
[19] ZHOU Z H, WU J X, TANG W. Ensembling neural networks: Many could be better than all[J]. Artificial Intelligence, 2002, 137(1/2): 239-263.
[20] ZHANG C X, ZHANG J S. A Survey of Selective Ensemble Learning Algorithms[J]. Chinese Journal of Computers, 2011, 34(8): 1399-1410.
[21] CHEN K. Study of selective ensemble alogrithm based on classification problems[J]. Application Research of Computers, 2009 (7): 2457-2459.
[22] ZHENG L R. Heuristic selective ensemble learning algorithm based on clustering and dynamic update[D]. Xiamen: Xiamen University, 2017.
[23] CHEN Q. Research on selective ensemble learning algorithm. Computer Technology and Development[J]. Comput Technol, 2010, 20(2): 87-89.
[24] FREY B J, DUECK D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.
[25] KUNCHEVA L I, WHITAKER C J. Ten measures of diversity in classifier ensembles: Limits for two classifiers[C]// Intelligent Sensor Processing. IET, 2001.
[26] YU J Y. Research on Enterprise Credit Risk Evaluation Based on Heterogeneous Learning Device Integration Strategy[D]. Beijing: Central University of Finance and Economics, 2019.
[27] LIU J C, JIANG X H, WU J P. Realization of a Knowledge Inference Rule Induction System[J]. Systems Engineering, 2003, 21(3): 108-110.
[28] LI Z S, LIU Z G. Feature selection algorithm based on XGBoost [J]. Journal on Communications, 2019(10).
[29] prosper-loan[EB/OL]. https://www.kaggle.com/yousuf28/prosper-loan.
[30] lendingclub[EB/OL]. https://www.lendingclub.com/info/download-data.action.

WANG Mao-guang, born in 1974, Ph.D, professor. His main research interests include intelligent risk control models and algorithms, big data and intelligent software engineering etc.

YANG Hang, born in 1997, postgraduate. Her main research interests include intelligent risk control models and algorithms.