Statistics and Application 统计学与应用, 2017, 6(4), 411-417
Published Online October 2017 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2017.64047

Hans汉斯

Personal Credit Assessment Model Based on Stacking Ensemble Learning Algorithm

Runze Peng

School of Mathematics and Systems Science, Beihang University, Beijing
Email: pengrunze@buaa.edu.cn

Received: Oct. 4th, 2017; accepted: Oct. 19th, 2017; published: Oct. 26th, 2017

Abstract

The prediction accuracy of traditional machine learning methods often depends on the specific problems. Ensemble learning achieves significant improvement in classification performance by combining several of base classifiers. This paper briefly introduces the basic idea of ensemble learning, discusses advantages of Stacking to the traditional classical ensemble algorithms. Based on the Stacking framework，we build two-layer classification model to evaluate the personal credit using the UCI datasets. The results of the empirical analysis show that, compared with the single machine learning method and simple average ensemble, Stacking with two-layer classifier has a better prediction effect.

Keywords

Ensemble Learning, Stacking, Credit Assessment

基于Stacking集成学习算法的个人信用评估模型

彭润泽

北京航空航天大学数学与系统科学学院，北京
Email: pengrunze@buaa.edu.cn

收稿日期：2017年10月4日；录用日期：2017年10月19日；发布日期：2017年10月26日

摘　要

传统机器学习算法的预测精度往往依赖于具体的问题，集成学习通过综合若干基分类器的预测结果，实

文章引用: 彭润泽. 基于 Stacking 集成学习算法的个人信用评估模型[J]. 统计学与应用, 2017, 6(4): 411-417.
DOI: 10.12677/sa.2017.64047

现了分类效果的显著提升。对集成学习的思想进行了简单地介绍，阐述了Stacking集成相对于传统经典集成算法的优势。并基于Stacking集成框架，利用UCI的信用评估数据集，构建两层分类器学习模型对个人信用进行评估。实证分析的结果表明，相对于单一的机器学习方法，以及对这些单一机器学习方法的结果进行简单的平均集成，两层分类器的Stacking集成学习有着更好的预测效果。

关键词

集成学习，Stacking，信用评估

1. 引言

随着我国经济进入稳步增长的新常态模式，国内金融信贷行业迅速发展，许多金融信贷服务逐步下沉到社会大众。而且受 2008 年次贷危机的影响，传统的企业金融业务增长乏力，大量金融机构开始愈加重视个人信贷业务，加上互联网金融的异军突起，个人的信用消费也变得越来越普遍，与之相辅相成的是对个人信用评估的迫切需求。

个人信用评估是基于个人的基本信息和历史信用数据，通过机器学习、数据挖掘等技术建立信用评分模型，对个人的信用进行量化评估。在个人信用评估研究领域，各种统计学方法和机器学习方法被大量运用[1] [2]。I-Cheng Yeh 和 Che-hui Lien 对 K 近邻、Logistic 回归，线性判别分析，朴素贝叶斯，人工神经网络，决策树这六种模型在信用卡违约概率预测问题上进行了比较，结果表明人工神经网络的预测准确度最高[3]。姜明辉等人针对单一模型在信用评估问题上存在的不足，将 RBF 神经网络和 Logistic 回归的预测结果进行加权组合，结果表明组合模型更加稳健且有着更高的预测精度[4]。叶晓枫和鲁亚会在随机森林特征选择算法的基础上构建朴素贝叶斯信用评估模型，发现两种算法组合后的模型有效提高了朴素贝叶斯模型的分类精度[5]。Paolo 的研究表明，结合马尔可夫链蒙特卡罗计算方法的贝叶斯模型，能够成功应用在基于高维度复杂数据集的信用评分问题上[6]。Lee 等人将反向传播神经网络与判别分析相结合，发现这种混合方法的收敛速度远快于传统的神经网络模型，而且信用评分的准确度要优于判别分析和 Logistic 回归[7]。Baesens 等人基于八个真实的信用评分数据集进行实证分析，发现最小二乘支持向量机和神经网络分类器预测性能非常优异，但是一些简单的分类器例如 Logistic 回归和线性判别分析同样表现良好[8] [9]。Farquad 等人提出一种联合 PCA 和 SVM 的信用评分模型，结果表明 PCA-SVM 混合模型的预测精度要高于单独的 SVM 模型[10]。David West 等人在多重感知机神经网络的基础上，利用 CV(交叉验证)，Bagging 和 Boosting 三种集成策略建立集成分类器，发现集成神经网络在降低泛化误差方面要显著优于最好的单一预测模型，但是三种不同集成策略的性能却没有明显的差别[11]。

本文基于 Stacking 集成算法的框架，构建了两层集成学习器，初级学习器采用四个预测精度较高的非参数机器学习算法，这些非参数方法对数据的分布不做假设，能够很好地捕捉大样本、高维度数据中的复杂非线性关系。次级学习器采用传统的统计学方法 Logistic 回归，Logistic 回归虽然分类精度低于一些非参数智能算法，但是简单易操作且有着很好的稳定性性，能够有效降低集成模型过拟合的风险。

2. 集成模型

集成学习通过对若干弱学习器进行适当的组合来得到预测性能较强的学习器。图 1 展现了集成学习

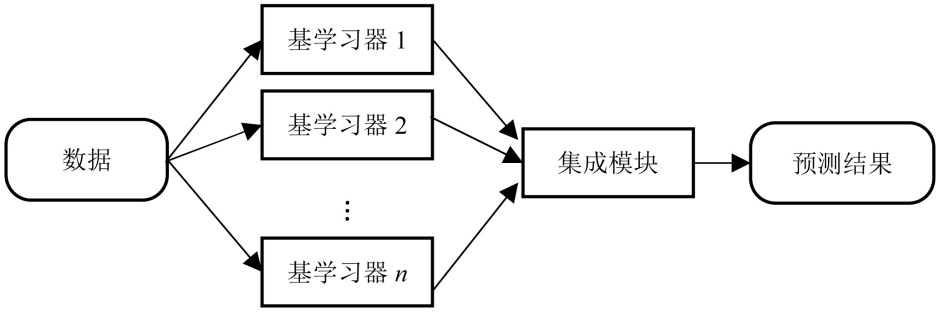

Figure 1. Structure of ensemble learning
图 1. 集成学习结构

的一般结构，每个基学习器由特定的学习算法从数据中产生，然后经过适当的组合策略来得到最终的预测模型。

集成学习预测效果的好坏主要取决于两个方面：一个是基分类器的预测精度，显然基分类器的分类准确率越高，集成学习的效果也会越好；另一个比较重要的方面就是基分类器的多样性，所谓多样性就是希望所有的基学习器相互之间能够有一定的差异。如果所有基学习器都产生了相同的预测结果，集成模型的预测效果也不会改善，反而会增加建模的复杂性。所以我们希望不同的基学习器能够“好而不同”[12]，从而实现不同基学习器之间的强强联合和优势互补。

Bagging 和 Boosting 是最具代表性的两种集成学习算法。Bagging 通过对给定的训练集进行等概率、有放回的抽样来得到若干训练集，然后用这些不同的训练集来训练出一系列基分类器。Boosting 在进行基分类器的构建时，会对之前预测错误的样本增加抽样权重，从而每一轮训练都是基于不同分布的数据样本，并在最后集成时会考虑每个基学习器的预测效果，进行加权平均。

Bagging 和 Boosting 一般都是基于单一的机器学习算法，比如决策树模型。而 Stacking 则是通过组合多种机器学习算法来提升分类器的泛化性能。它先从原始训练集中训练出初级学习器，然后以初级学习器的预测结果作为特征来训练一个次级模型，比如最简单的次级模型就是对多个初级基学习器的结果进行简单投票。Stacking 依靠不同学习算法的差异来保证基学习器的多样性，而且通过次级学习器以最佳的方式来整合不同基学习器的预测结果，相对 Bagging 和 Boosting，Stacking 往往预测精度更高，而且过拟合的风险会更低[13]。

本文构建的个人信用评估模型的示意图如图 2 所示。

我们先对原始数据采取一些预处理措施和特征降维，然后将数据划分成训练集和测试集两部分，分别进行模型的训练和结果的预测。为了便于比较 Stacking 和 Bagging、Boosting 的预测效果，我们将 Bagging 的代表性算法 RF (随机森林)和 Boosting 的代表性算法 GBDT (梯度提升树)作为初级学习器的组成部分，另外两个初级的基学习器是经典的机器学习分类算法 SVM (支持向量机)和 ANN (人工神经网络)，次级学习器采用稳定性较高的 Logistic 回归算法。

3. 实证分析

本文采用来自 UCI 的台湾地区信用卡客户违约记录数据，包含 30,000 条样本，24 个变量，具体的变量描述如表 1 所示，响应变量为 *Y*，表示客户是否违约，其余 23 个变量都是解释变量，其中包含 9 个离散变量，14 个连续变量。

3.1. 数据预处理和降维

我们对客户的年龄变量采取离散化处理，将客户年龄划分成 5 个区间：25 岁以下，25~35 岁，35~45

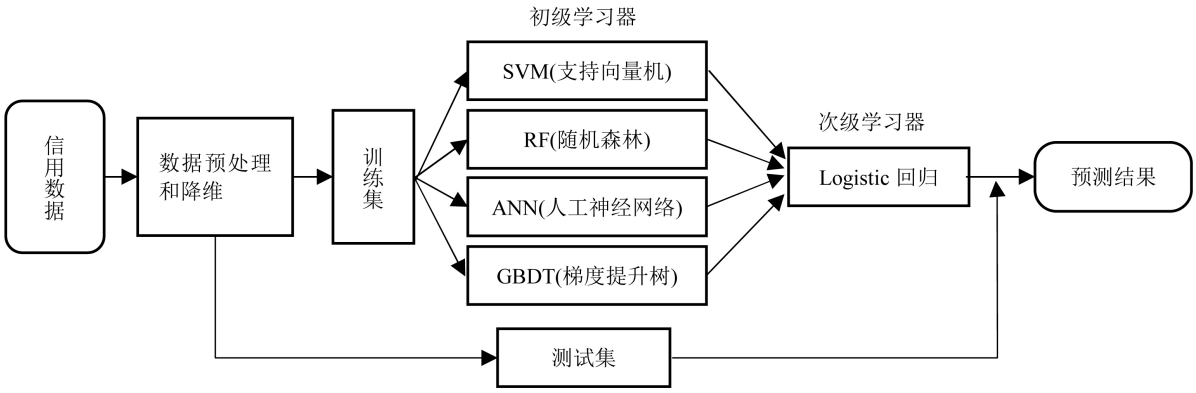

Figure 2. Structure of personal credit assessment model
图 2. 信用评估模型结构

Table 1. Description of variables
表 1. 变量描述

变量名称	变量含义
Y	是否违约：1=违约，0=未违约
X_1	银行给予客户的信用额度(包括个人信用额度和客户的家庭信用额度)
X_2	客户的性别：1=男性，2=女性
X_3	客户的教育水平：1=研究生及以上，2=大学，3=高中，4=其它
X_4	客户的婚姻状况：1=已婚，2=未婚，3=其它
X_5	客户的年龄
$X_6 - X_{11}$	这 6 个变量依次表示客户从 2005 年 4 月到 9 月每月的还款情况：−1=及时还款，1=还款延迟一个月，2=还款延迟两个月，…，9=还款延迟九个月及以上
$X_{12} - X_{17}$	这 6 个变量依次表示客户从 2005 年 4 月到 9 月每月的信用卡消费金额
$X_{18} - X_{23}$	这 6 个变量依次表示客户从 2005 年 4 月到 9 月每月的支付金额

岁，45~60 岁，60 岁以上。变量离散化后对异常数据有更强的鲁棒性，并且能够提升学习算法的训练效率。另外我们对数据采用如下公式所示的标准化处理：

$$X^* = \frac{\left(X - \bar{X}\right)}{3\sigma} \tag{1}$$

其中 $\bar{X}$ 表示变量的均值，σ 表示变量的标准差。根据正态分布的 3σ 原理，标准化后大概有 99.7% 的变量值被压缩到−1 和 1 之间，对于落在区间[−1,1]之外的值均设为−1 和 1，这样处理既能保证变量的取值全都落在一个相同且较小的范围内，消除不同变量量纲和数值的差异，又能降低一些极端值对建模带来的负面影响。

图 3 是 13 个连续型解释变量的相关系数图，图中数字颜色的深浅反映出相关系数绝对值的大小，从图 3 可以明显看出 $X_{12} - X_{17}$ 这 6 个解释变量呈现出非常强的正相关性，存在较多重叠和冗余的信息。

为了剔除这几个变量中存在的大量冗余信息，降低建模的复杂度，我们采用主成分分析方法对这几个解释变量进行降维处理。主成分分析通过对协方差矩阵做奇异值分解，将大量相关性很高的变量转化成几个相互独立、且能解释大部分原始数据信息的主成分。主成分分析的结果如表 2 所示，从表中可以看到第一主成分就贡献了原始数据 90%的信息，因此我们用第一主成分代替 $X_{12} - X_{17}$ 这六个解释变量。

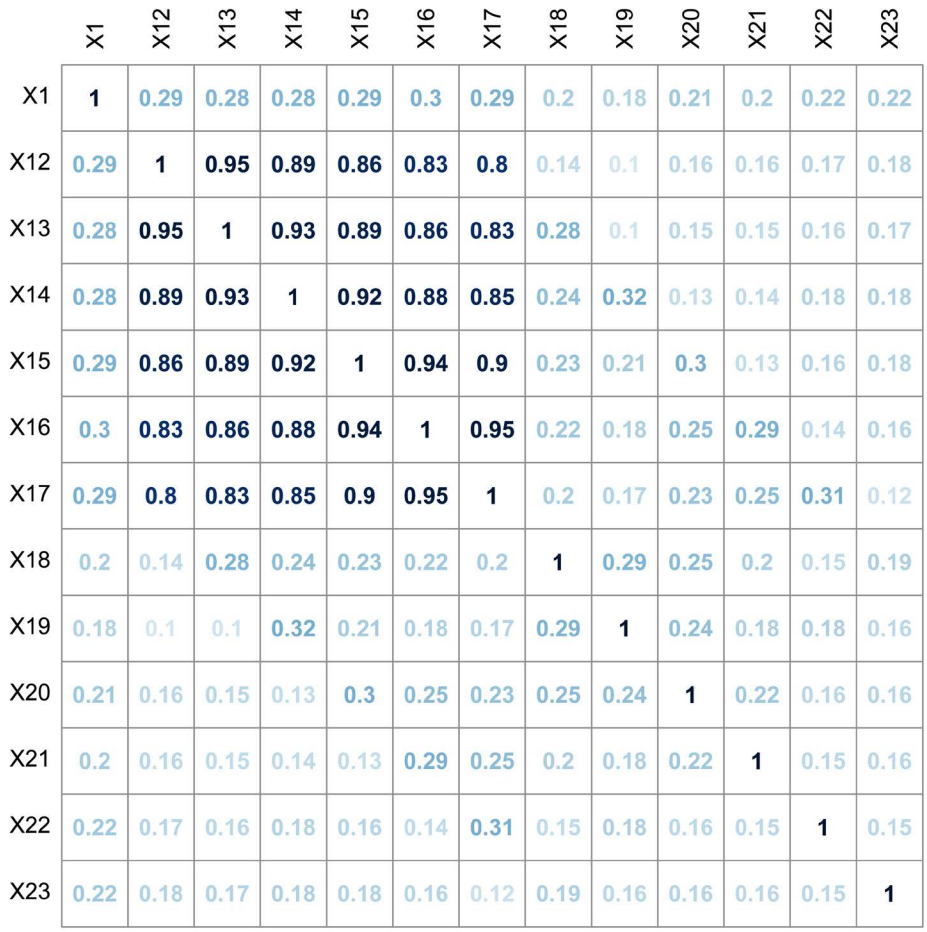

Figure 3. Correlation coefficient
图 3. 相关系数

Table 2. Results of principal component analysis
表 2. 主成分分析结果

主成分	1	2	3	4	5	6
特征值	5.433	0.306	0.112	0.067	0.042	0.040
贡献率%	90.550	5.099	1.861	1.119	0.693	0.678
累积贡献率%	90.550	95.649	97.510	98.629	99.322	100

3.2. 预测分析和结果比较

我们利用分层随机抽样按照 3:1 的比例将原始数据划分成训练集和测试集两部分，训练集包含 22,500 个样本，测试集包含 7500 个样本。模型的建立主要利用 R 软件中的 caret 包来完成。对于所有的基学习器，均采用 10 折交叉验证的网格搜索法来确定重要超参数的取值。选择分类正确率和 AUC 值作为模型的评价指标。分类正确率是正确分类的样本数占总样本数的比例，反映了模型的整体分类精度。AUC 值代表模型 ROC 曲线下的面积，常被用来判断一个二值分类器的优劣，是对正例分类精度和反例分类精度的综合度量，AUC 取值在 0.5 到 1 之间，越接近 1 表示分类器的效果越好[14]。所有模型的预测结果如表 3 所示。

Table 3. Prediction results of different models
表 3. 不同模型预测结果

模型	分类正确率		AUC 值	
	训练集	测试集	训练集	测试集
SVM (支持向量机)	81.84%	81.76%	0.7830	0.7303
RF (随机森林)	99.32%	82.12%	0.9995	0.7741
ANN (人工神经网络)	83.31%	80.85%	0.8152	0.7583
GBDT (梯度提升树)	83.22%	82.29%	0.8191	0.7885
简单投票法	91.20%	82.24%	0.9616	0.7873
Stacking 集成模型	94.66%	85.31%	0.9711	0.7901

在表 3 的 6 个模型当中，前 4 个模型是我们建立集成模型所采用的四个单一初级分类器，第 5 个模型简单投票法采用“软投票”策略，即对所有初级分类器的输出概率求算术平均，然后选择合适的阈值进行结果预测。最后一个模型就是我们基于 Stacking 构建的两层集成模型，在这个模型当中，同样是利用初级学习器的输出概率来训练次级学习器，相比于初级分类器产生的二值分类结果，概率结果包含了更多基分类器在数据里发现的模式和信息。

由表 3 可知在训练集上 RF 的分类正确率最高，几乎达到了百分之百，简单投票法和 Stacking 集成模型的分类正确率也都超过了 90%，另外三个模型的分类正确率都低于 85%。在更能反映模型泛化能力的测试集上，Stacking 集成模型的分类正确率最高，达到了 85.31%，比最好的单一分类器 GBDT 高了 3.02%，而在训练集上表现非常好的 RF 的分类正确率只有 82.12%，这说明 RF 存在着一定程度的过拟合。简单投票法在训练集和预测集上的分类正确率分别达到了 91.20%和 82.24%，要高于四个单一模型 86.92%和 81.76%的平均分类正确率，这说明简单的集成也能在一定程度上提升分类效果，但简单投票法在训练集和测试集的分类正确率都要显著低于 Stacking 集成模型，表明通过训练第二层的 Logistic 分类器能够产生对初级学习器更好的组合方式。AUC 值的结果同样体现了 Stacking 集成模型有着最优的泛化性能。

4. 结束语

近年来，随着我国经济的快速发展，消费信贷业务也增长迅速，住房按揭、助学贷款、信用卡等各种个人信贷业务的规模不断增长。但是相对于西方发达国家，我国个人征信系统的建设并不完善，大量人群的信用数据严重缺失，很多金融机构开始广泛采集用户的基本资料、信贷记录、信用卡使用情况等各种信息来构建自身的数据仓库。但是这些数据往往存在着样本量大、维度高、多元化、冗余化等特点，这也对传统机器学习模型的处理能力提出了挑战。集成学习的出现很好地解决了这一问题，其通过产生若干基学习器然后进行组合，对原始数据的信息进行了最大化利用，获得了比单一学习器更加优越的泛化性能。

本文利用 Stacking 集成框架，构建了两层分类器的个人信用评估模型，第一层模型采用分类精度较高的机器学习算法，第二层模型采用稳健性较好的 Logistic 回归方法，实现了预测精度和模型稳健性的统一。在 UCI 信用数据集上的实证分析结果表明，Stacking 集成模型的分类精度要优于单一的机器学习模型，包括基于 Bagging 和 Boosting 构建的单一模型，同时该模型的效果也要优于仅仅对多个单一模型进行简单投票表决的方法，这为金融机构建立个人信用评估系统提供了一种新的思路和方法，也为构建多层 Stacking 集成的个人信用评估模型提供了参考和借鉴。

参考文献 (References)

[1] Hand, D.J. and Henley, W.E. (1997) Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society*: *Series A* (*Statistics in Society*), **160**, 523-541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

[2] García, V., Marqués, A.I. and Sánchez, J.S. (2012) Non-Parametric Statistical Analysis of Machine Learning Methods for Credit Scoring. *Management Intelligent Systems*, 263-272. https://doi.org/10.1007/978-3-642-30864-2_25

[3] Yeh, I.C. and Lien, C.H. (2009) The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. *Expert Systems with Applications*, **36**, 2473-2480. https://doi.org/10.1016/j.eswa.2007.12.020

[4] 姜明辉, 谢行恒, 王树林, 等. 个人信用评估的 Logistic-RBF 组合模型[J]. 哈尔滨工业大学学报, 2007, 39(7): 1128-1130.

[5] 叶晓枫, 鲁亚会. 基于随机森林融合朴素贝叶斯的信用评估模型[J]. 数学的实践与认识, 2017(2): 68-73.

[6] Giudici, P. (2001) Bayesian Data Mining, with Application to Benchmarking and Credit Scoring. *Applied Stochastic Models in Business & Industry*, **17**, 69-81. https://doi.org/10.1002/asmb.425

[7] Lee, T.S., Chiu, C.C., Lu, C.J., *et al*. (2002) Credit Scoring Using the Hybrid Neural Discriminant Technique. *Expert Systems with Applications*, **23**, 245-254. https://doi.org/10.1016/S0957-4174(02)00044-1

[8] Baesens, B. (2003) Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation. *Management Science*, **49**, 312-329. https://doi.org/10.1287/mnsc.49.3.312.12739

[9] Stepanova, M. (2003) Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *Journal of the Operational Research Society*, **54**, 627-635. https://doi.org/10.1057/palgrave.jors.2601545

[10] Farquad, M.A., Ravi, H., Sriramjee, V., *et al*. (2011) Credit Scoring Using PCA-SVM Hybrid Model. *Communications in Computer & Information Science*, **142**, 249-253. https://doi.org/10.1007/978-3-642-19542-6_40

[11] West, D., Dellana, S. and Qian, J. (2005) Neural Network Ensemble Strategies for Financial Decision Applications. *Computers & Operations Research*, **32**, 2543-2559. https://doi.org/10.1016/j.cor.2004.03.017

[12] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 171-173.

[13] Zenko, B., Todorovski, L. and Dzeroski, S. (2001) A Comparison of Stacking with Meta Decision Trees to Bagging, Boosting, and Stacking with other Methods. *IEEE International Conference on Data Mining*, 669-670. https://doi.org/10.1109/ICDM.2001.989601

[14] Fawcett, T. (2006) An Introduction to ROC Analysis. *Pattern Recognition Letters*, **27**, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010