文章编号: 1000-2243(2008)S0-S153-04

基于随机森林的企业信用评估模型

彭国兰[1], 林成德[2]

(1. 厦门城市职业学院工程技术学部, 福建 厦门 361008; 2. 厦门大学自动化系, 福建 厦门 361005)

摘要: 引入了一种能较好容忍噪声, 且稳定性较高的组合分类器算法——随机森林(RF), 建立企业信用评估模型; 着重分析了适合 RF的不平衡分类问题的处理方法, 并介绍了模型参数的优化. 通过与神经网络和支持向量机的对比实验, 证实了该方法的有效性和优越性.

关键词: 信用评估; 随机森林; 不平衡分类.

中图分类号: TP18 文献标识码: A

A model based on random forests for enterprises credit assessment

PENG Guo-lan[1], LIN Cheng-de[2]

(1. Department of Engineering Technology, Xiamen City University, Xiamen, Fujian 361008, China;
2. Department of Automation, Xiamen University, Xiamen, Fujian 361005, China)

Abstract We introduce a new classifier combination algorithm——random forests (RF), which is rather stable and robust with noise. By analyzing the real data, a new model based on RF is built and tested. Empirical results show that the new proposed model is effective and more advantageous than those of both neural network model and SVM model.

Keywords credit assessment; random forests; unbalanced classification

企业信用评估是商业银行资产业务, 特别是贷款业务经营的核心内容, 主要是对贷款企业的信用风险进行评价. 从技术角度看, 企业信用评估实质是一个分类问题, 即根据相关评价指标将企业划分为不同的信用等级. 进入 90年代以来, 人工神经网络(ANN)和支持向量机(SVM)等智能评估模型已被引入到信用评估领域, 并取得了令人鼓舞的成果. 但前人的研究着重于如何提高单分类器模型的性能, 并且评估指标的确定大多采用专家建议的财务比率指标, 带有一定的主观性; 另外, 由于企业信用评估本身特点, 其样本数据分布复杂且噪声较多, 这使得单分类模型难以达到令人满足的效果. 近年来, 多分类器组合算法逐渐成为智能算法研究的一个主要分支. 随机森林是一个组合分类器算法, 该算法能较好地容忍噪声, 且稳定性较高. 本研究综合考虑随机森林的特点, 探索研究随机森林在企业信用评估领域中的应用.

为提高评估指标的客观性, 确定适合智能评估模型的指标体系, 我们将随机森林技术引进到企业信用评估领域的应用, 得到了更为客观、且适合随机森林建模的指标体系[1]. 文[1]的工作主要集中于随机森林在特征选择中的应用, 对于如何调整模型参数以优化模型性能涉及较少. 本文在文[1]的基础上重点研究了随机森林模型参数的优化, 进一步提高了模型的性能, 提出了适合 RF模型的不平衡分类问题的处理方法, 以平衡各等级企业的评价准确率; 并经由仿真实验优化了模型参数. 通过与 ANN和 SVM的对比实验, 证明了 RF模型的优越性.

1 随机森林简介

随机森林[2] (random forests, RF)是 Breiman于 2001年提出的一种新的组合分类器算法, 采用分类回归树(CART)[3]作为元分类器, 用 Bagging方法制造有差异的训练样本集, 并且在构建单棵树时, 随机地选择特征对内部节点进行属性分裂. Bagging方法和 CART算法的结合, 再加上随机选择特征进行属性分

收稿日期: 2008-06-13

作者简介: 彭国兰(1976-), 女, 硕士研究生, 助教; 通讯联系人: 林成德, 教授.

裂, 使得 RF能较好容忍噪声, 并具有较好的分类性能. Breiman通过理论证明了随机森林算法存在误差上界.

2 基于 RF的企业信用评估模型

2 1 样本预处理

实验数据来源于福建省某商业银行 2003年的贷款企业数据库, 共有 1 282家企业的相关数据, 其中信用等级[4]为 AAA的企业有 650家, 等级为 AA的企业有 518家, 等级为 ABC①的企业有 114家.

2 1 1 样本的野点删除

随机森林将野点定义为: 与数据集中其它所有样本点的相似度都很小的样本点, 样本相似度为原始训练集中两两样本之间的相近程度. 另一个更为实用的定义是: 与类别 j中其它所有样本点的相似度都很小的样本点为类别 j中的野点.

利用随机森林的野点检测方法[5], 计算出各样本点的野点度量值, 然后确定一个阈值, 删除野点度量值大于阈值的样本点. 删除野点后, 新的样本数据共 1 250条, 其中: 信用等级为 AAA的企业 649家, AA的企业 499家, ABC的企业 102家; 用该数据集作为本文的实验数据.

2 1 2 特征选择

1)候选指标全集. 根据常用的准则以及专家的建议, 考虑随机森林可以处理离散指标的特点, 最终采用文[1]中的 24个财务比率指标及 "企业和领导者素质", 共 25个指标作为评估模型的候选指标全集. 前 24个指标标号与文[1]一致, 新增指标 "企业和领导者素质"为 25号指标.

2)特征选择. 随机选取总体数据的 75% 作为训练数据, 剩余的 25% 留作测试数据. 使用训练数据的五重交叉数据选取特征子集[1], 最终得到的特征子集为{3, 5, 6, 8, 9, 10, 13, 14, 19, 21, 24, 25}, 共 12个指标.

2 2 模型的构造

实验主要分为 3部分: ①确定适合 RF的不平衡分类问题的处理方法; ②优化 RF模型的参数; ③建立三等级评估模型.

2 2 1 RF对不平衡分类问题的处理方法

企业信用评估问题, AAA和 AA类的数据相对较多, ABC类数据较少, 这样会导致算法偏向正类样本(本文指 AAA和 AA类样本), 而使负类样本(ABC类)的预测准确率很低. 这意味着把信用度低的企业错判为信用高的企业比率增大, 这样将给银行带来巨大的风险, 甚至让银行面临本金无法收回的巨大损失. 于是应该采取相应的处理方法, 尽量平衡各类的分类准确率. 陈超在文[6]中指出, RF处理不平衡分类问题有 2大方法: ①基于采样技术的方法; ②基于代价敏感学习的方法.

建模过程中分别使用这 2种方法处理不平衡分类问题, 对比实验结果见表 1.

从表 1中可以看出, 用权重法处理不平衡分类问题时, 在保持较高分类准确率的基础上, 同时均衡了各类的准确率. 因此, 本文在建模过程中采用权重法处理不平衡分类问题.

表 1 处理不平衡分类问题的方法比较

Tab. 1 Comparison of methods on dealing with unbalanced classification

处理方法	整体预测误差率 /%	AAA /%	AA /%	ABC /%
不做任何处理	12.87	10.17	12.61	32.00
等量抽样法	23.32	20.90	27.93	20.00
权重法② Classwt= [1 1 3 3.5]	14.38	13.56	16.32	16.00

① 由于得到的评估数据中 A、B、C 3个等级的数据太少, 因此本文将这 3个等级进行合并, 记为 ABC.

② 权重法中的 Classwt为实验采用的权重向量值.

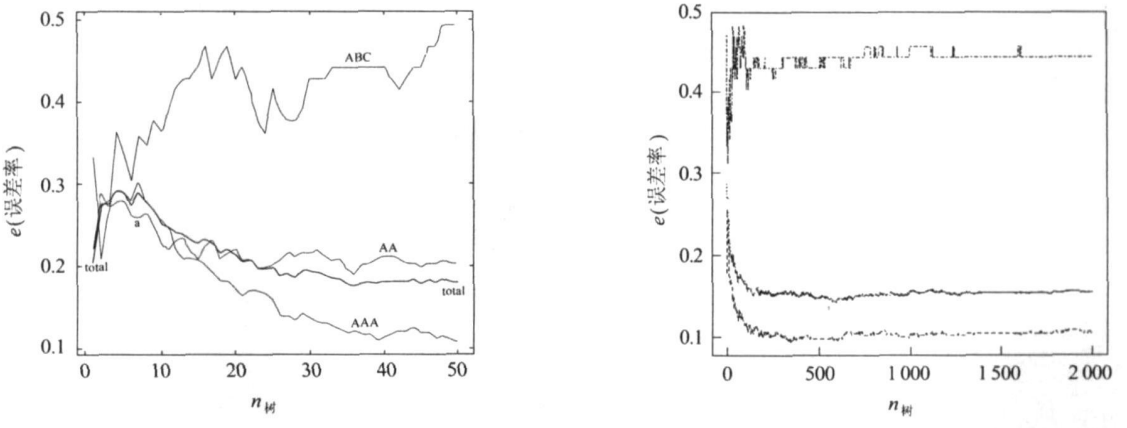
2 2 2 RF模型参数的选择

RF模型有 2个比较重要的可调参数：①ntree森林中树的数目；②mtry每个节点处候选特征的个数. ntree的设置相对较简单，只要让 ntree的值足够大，以保证 RF收敛即可.

通过实验考察 ntree和 mtry对模型性能的影响，并确定参数的取值.

1) mtry取缺省值 sqrt(M)，令 mtry= 3，ntree分别取值为 (50，2 000)，建立 RF模型. 实验结果如图 1所示，横轴为随机森林中树的个数的取值，纵轴为误差率 (%).

从图 1(a)可以看到 RF模型没有达到最佳性能，随着 RF中树的数目的增长，模型的整体误差率及各类误差率都没有趋于稳定，这表明 RF模型的 ntree的取值不够大，应加大 ntree的取值. 而从图 1(b)可以看到，整体误差率和各类误差率基本趋于稳定，于是选定模型的 ntree= 2 000.



(a) n_{tree} = 50　　(b) n_{tree} = 2 000

图 1　ntree 的不同取值对 RF 模型性能的影响

Fig. 1　Performance of model on different value of ntree

2) 参数 mtry为在各节点处候选的特征个数，mtry是 RF唯一的一个较敏感参数，调整 mtry的取值，模型的准确率变化较为明显. mtry的缺省设置为 sqrt(M).

在仿真实验中，固定的取值，调节的取值，观察误差的变化. 利用上文实验结果，令 ntree= 2 000，取值为 (1，2，…，M)，模型的误差随 mtry的变化如图 2所示.

从图 2可以看到，当 mtry= 1时，模型的误差率取得最小值，因此选定模型的 mtry的值为 1.

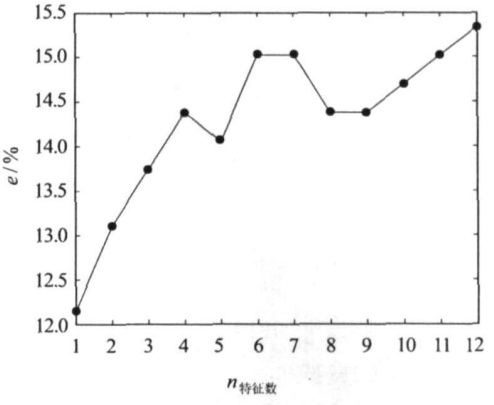


图 2　模型误差随 mtry的变化

Fig 2　Error rate on different value of mtry

2 2 3 评估模型的建立

综合以上实验结果，利用 RF建立三等级评估模型，并利用 SVM 和 NN的各种方法建立三等级信用评估模型进行对比实验，实验结果见表 2

3 结果分析

从表 2的对比结果可以看出，在 SVM和 NN的各模型中，DAG SVM模型取得的最高整体预测准确率

为 83.15%,其中 AAA类的准确率为 86.81%, AA类为 80.70%, ABC类为 71.43%,类别的最高准确率与最低准确率相差 15.38%.虽然 SVM可以通过调整各类别的惩罚系数以达到平衡各类准确率的目的,但平衡类间准确率会导致整体准确率的下降.而 RF模型的预测准确率达到了 87.86%,其中 AAA类的准确率为 88.70%, AA类为 86.49%, ABC类为 88.00%,类别最好预测准确率与最差准确率仅相差 2.21%;即 RF模型在平衡了各类准确率的基础上,仍能达到较高的整体准确率. RF模型的整体准确率比 DAG SVM模型的准确率高出 4.71%.表 2的结果表明, RF用来建立评估模型时,比以往方法(SVM, NN)表现出更优的性能.

表 2 三等级评估结果的比较

Tab.2 Comparison of performance of model on different methods

建模方法	训练准确率/%				测试准确率/%			
	整体	AAA	AA	ABC	整体	AAA	AA	ABC
1-vs-1 SVM	92.83	92.59	93.27	92.06	82.80	87.50	79.82	66.67
DAG SVM	92.95	92.59	93.56	92.06	83.15	86.81	80.70	71.43
1-vs-n SVM	90.92	92.59	90.64	80.95	78.85	86.81	75.44	42.86
ECOC SVM	89.73	95.37	86.55	68.25	78.85	94.44	70.18	19.04
NN平均	92.13	95.56	87.57	93.33	78.03	92.22	61.67	69.52
NN最优	92.11	95.60	88.30	88.89	81.36	93.75	67.54	71.43
RF	83.99	84.75	82.47	87.01	87.86	88.70	86.49	88.00

注: RF模型的训练准确率为原始训练数据集的 OOB准确率

4 结语

首先利用文[1]中特征选择的方法得到企业信用评估的指标体系,然后着重介绍了如何建立 RF评估模型,分析得到了适合 RF的不平衡分类问题的处理方法,介绍了如何设置模型参数.通过仿真实验比较了 RF模型与 SVM和 NN模型的性能,结果表明,随机森林用于企业信用评估模型的特征选择,并建立企业信用评估模型时,能达到更好的性能.

通过在真实数据集上的仿真实验,证实了 RF较优的性能和较高的鲁棒性,具有较好的发展前景,值得深入研究.未来的工作可以从以下方面深入:

1)采用 RF进行特征选择,得到的指标更加客观且符合模型特点,可以考虑扩大候选指标集,尽可能提取数据中的有用信息.

2)使用 RF删除野点时,野点阈值的确定值得进一步深入研究.

3)对 RF模型参数的选取有了比较清晰的思路,值得进一步深入研究.

4) RF算法是利用所有树参与最终决策,如何选择那些更有效的树进行最终决策值得研究.

参考文献:

[1] 林成德,彭国兰.随机森林在企业信用评估指标体系确定中的应用[J].厦门大学学报:自然科学版, 2007, 46(2): 199-203.

[2] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[3] Breiman L, Friedman J, Olshen R, *et al*. Classification and regression trees[M]. New York: Chapman & Hall, 1984.

[4] 何艳芳,石丹林.银行客户信用评估[M].北京:中国商业出版社, 2002.

[5] 邱一卉,林成德.基于随机森林方法的异常样本检测方法[J].福建工程学院学报, 2007(4): 392-396.

[6] Chao Chen, Andy Liaw, Leo Breiman. Using random forest to learn imbalanced data[J/OL]. http://www.stat.berkeley.edu/users/chenchao/666.pdf

(责任编辑:王阿军)