基于组合分类器的上市公司信用风险评价

张 杰 王 凡

(北京工业大学经济与管理学院 北京 100022)

【摘要】针对传统信用风险评价模型只含有一个分类器的缺陷,本文利用 AdaBoost 组合分类器来对上市公司信用风险进行评价,并与基于支持向量机和神经网络的分类模型进行了效果比较。实证研究表明,组合分类器克服了单一分类器的诸多缺点,预测准确率高于单一分类器。

【关键词】组合分类器 AdaBoost 信用风险评价

一、信用风险及判别方法简述

信用风险是指在以信用关系为纽带的交易过程中,交易一方不能履行给付承诺而给另一方造成损失的可能性。自 Altman 在企业财务危机及信用风险分析方面所做的开创性工作以来,多元统计分析特别是多元判别分析技术(MDA)获得了广泛应用。MDA 的最大优点在于其具有较好的解释性和简明性,其缺陷是设定有较严格的前提条件——要求数据服从多元正态分布和协方差矩阵相等。国外学者对 MDA 的缺陷从不同角度进行了改进,形成了两类模型,即统计模型和人工智能模型。

在统计模型方面,以 Altman 提出的 Z 值模型为代表,以财务信息作为输入变量,运用多元统计方法得到与企业信用显著相关的变量组合。Z 值模型直观可行,操作简便,但由于它的精确性对企业所处的行业、所在的国家和所处的历史时期是非常敏感的,所以其在各国间的应用有相当大的差异,直接使用该模型的效果并不理想。在人工智能模型方面,神经网络技术(NN)在 20 世纪 90 年代被引入企业信用评价中,NN 是一种对数据分布无任何要求的非线性技术,能有效解决非正态分布、非线性的预测评估问题,但它存在结构确定难度大、训练样本集大和训练效率低等缺点。

以上模型对评价我国企业的信用风险都有一定的参考价值,但是单一分类技术在应用中常常会受到一定条件的限制。因此,需要研究更好的分类模型来不断提高上市公司信用风险评价模型的判别精度。传统的分类学习框架都是通过学习构造出一个分类器,期望能够对未知数据实现最佳拟合。与此不同,组合分类器利用多个个体学习器解决同一个问题,它克服了单一分类器的诸多缺点(如对样本的敏感性、分类精度难以提高等等),已经在字符识别、文本分类、面部表情识别等领域获得了较好的应用效果。本文将 AdaBoost 组合分类器应用到上市公司信用风险评价中,提高了分类正确率。

二、AdaBoost 组合分类器

1. 组合分类器。所谓组合分类器是指几个分类器通过某种策略组合在一起进行分类。组合的策略可以是模型组合,可以是不同的算法组合,也可以通过对样本取样、变化样本空间、构造不同的分类器,然后按照一定的加权方法对分类器进行组合,最后确定可用的分类器。1995 年,Freund 和 Schapire 提出了 AdaBoost 算法。AdaBoost 即自适应提升算法,解决了早期 Boosting 算法很多实践上的困难,不需要预先知道弱学习器学习正确率的下限,可以很容易应用到实际问题中。由于 AdaBoost 算法最后结果的准确度依赖于弱学习返回的所有假设,而不是只依赖于准确率最低的那个假设,因此它可以全面开发弱学习的能力。

2. AdaBoost 组合分类器算法及分析。AdaBoost 算法的基本过程是:依次训练一组分量分类器,其中每个分量分类器的训练集都是选择由其他分量分类器给出的"最富信息"的样本组成,最后用线性加权集成这些分量分类器,从而得出最终判断结果。其中,"最富信息"样本的选取方法为:每个训练样本都被赋予一个权重,表明它被某个分量分类器选入训练集的概率。如果某个样本被当前弱分类器准确分类,那么它的权重就会降低,则在构造下一个分量分类器的训练集时,它被选中的概率会降低;相反,如果某个样本没有被正确分类,那么它的权重就会相应提高,它入选下一个分量分类器的训练集的概率也会提高。通过这种方式,AdaBoost 能够"聚焦于"那些比较容易出现分错的样本。

在具体实践上,令每个训练样本的初始权重相等,对于第 t 次迭代操作,需要根据第(t-1)次训练得到的样本权重来选取新的训练样本集,进而训练分类器 C_t。然后,用分类器 C_t 对整个样本集进行测试,提高被它错分的样本的权重,同时降低可以被正确分类样本的权重。之后,权重更新过的样本集被用来训练下一个分类器 C_{t+1},整个训练过程如此迭代进行,直到满足结束条件为止。其具体分类过程如下:

假设样本集合为$[(x_1,y_1),\cdots,(x_m,y_m)]$;$x_i \in X$ 为原始样本,$y_i \in \{-1,1\}$为类别标号。那么:

第一步:初始化权值 $D_1(i)=1/m$,m 为样本个数。

第二步:For t=1,…,T。

(1)权值归一化。

(2)对每一个特征 j,训练一个弱学习器 h_j,学习器的错误率用样本分布的权值 W_t 来衡量:

$E_j=\Sigma_i W_t|h_j(X_i)-Y_i|$

(3)选出一个错误率最低的弱学习器 h_t,其错误率为 E_t。

(4)更新权值: $W_{t+1,i}=W_{t,i}\beta_t^{1-e_i}$,其中当 h_t 对样本分类正确时,$e_i=0$,否则 $e_i=1$,$\beta_t=E_t/1-E_t$。

(5)最终强分类是:$h(x)=\begin{cases}1, \sum_{t=1}^{T}\alpha_t h_t(x)\geqslant\frac{1}{2}\sum_{t=1}^{T}\alpha_t\\ 0,\text{其他}\end{cases}$,

其中 $6\alpha_t=\log(1/\beta_t)$。

在 AdaBoost 组合分类器方法中,增加错分样本的权值并减少正确分类样本权值的结果是更高权值的样本对训练中的分类器影响更大,因此使分类器更关注错分样本,这些错分样本通常是最靠近决策边界的样本。

在多数情况下,每个分类器只要是弱分类器,即分类准确率超过 50%,会比随机猜想的要好,那么组合分类器的训练误差就随着 T 的增大而变得很小。同时组合分类器能保持良好的泛化能力,即使在 T 很大的情况下也很少会出现过拟合现象。在弱分类器的选择上,必须考虑算法的分类精度、稳定性和泛化能力等多个方面。对于不稳定的弱分类器,使用 AdaBoost 算法可以改善其分类准确率。如果分类器是稳定的,即训练数据集中的变化只在分类器上引起很小的变化,则 AdaBoost 对性能改善作出的贡献通常将很小。

三、基于组合分类器的上市公司信用风险实证分析

1. 上市公司样本选取及数据处理。证监会规定上市公司连续两年亏损即被特别处理(ST),可见 ST 公司的财务状况是不好的,它比一般的上市公司存在着更大的信用风险。因此,本文将因财务状况异常而被特别处理的上市公司界定为陷入财务困境的公司,同时将与之对应的同样数量的近期不会发生财务困境的公司作为经营状况好的样本组。将总样本分为两组:一组为训练样本组,用来构建预测模型;另一组为测试样本组,用来测试预测模型的分类准确率。

本文选取沪深两市 2004 年的 40 家上市公司(ST 公司 20 个和非 ST 公司 20 个)和 2005 年的 54 家上市公司(ST 公司 27 个和非 ST 公司 27 个)作为总样本。其中处于财务困境的上市公司样本和经营状况好的上市公司样本各为 47 家。根据上市公司信息披露制度的规定,上市公司必须对披露信息的真实性负责,因此上市公司前一年的财务数据可以反映其当年的信用状况。本文选取的数据是 ST 公司被宣布特别处理前一年的财务数据,非 ST 公司是按照其所对应的 ST 公司选取财务数据年份选取数据的。其中将所选的 2004 年的 40 家上市公司作为训练样本,所选的 2005 年的 54 家上市公司作为测试样本。这样共有 40 个训练样本和 54 个测试样本。

2. 指标选择。为了提高模型的预测能力,确保入选模型解释变量的每一个指标都具有显著的预测能力,本文从 Beaver(1966)、Altman(1968)、吴世农和卢贤义(2003)、陈晓(2003)、章之旺和吴世农(2005)等国内外学者有关财务困境预警研究的文献中挑选出经原作者实证检验预测能力显著的指标,添加笔者根据理论分析认为预测能力显著的其他指标,经整理形成分别反映公司盈利能力、偿债能力、营运能力、成长能力和现金流量 5 个方面共 8 个指标。具体包括偿债能力指标:流动比率、负债比率;营运能力指标:存货周转率、总资产周转率;盈利能力指标:净资产收益率、每股收益;成长能力指标:总利润增长率;现金流量指标:每股经营现金流量(所有指标均来自证券之星公布的沪深上市公司财务综合指标)。这些指标的选用,既考虑了公司的资产与负债能力,又兼顾到公司的盈利能力与成长能力,能够充分体现公司的信用状况。

3. 模型及参数的选择。AdaBoost 组合分类器用于单分类器的组合训练和结果融合,需要一个分类算法作为它的弱分类器。由于差异性是影响组合分类器泛化性能的重要因素,而 AdaBoost 分量分类器的精度和它的差异性又互为矛盾,即 AdaBoost 两个分量分类器的精度越高,它们之间的差异就越小,因而只有当精度和差异性达到某种平衡时,AdaBoost 才能体现出较好的性能。

神经网络技术虽然有很好的非线性拟合能力,但它存在训练速度慢、易陷入局部极小点、泛化能力差、网络结构和初始权值难以确定等缺点。RBFSVM 有高斯宽度 σ 和规则化参数 C 两个参数,任一个参数的改变都会导致分类器性能的改变。通过选择合适的 C 和 σ可以避免出现过拟合情况,即:若 C 值过小,则分类器学习能力不好,而当 C 在一个合适的范围内取值时,RBFSVM 的性能可以简单地通过调整值改变,且对分类器的影响更大。本文选取的 40 个训练样本和 54 个测试样本,采用 RBFSVM 作为弱分类器,分类器数量 T 取 10,惩罚参数 C=1 000,样本权重实现方式采用 Resampling。

4. 分类结果及分析。为了考察组合分类器的实际分类效果,本文用同样的训练数据和测试数据对基于 AdaBoost、SVM 和神经网络分类算法作了实证分析,分类结果如下:

AdaBoost、SVM、BP神经网络的Logistic回归分类结果

模　型	分类正确率(%)	
	训练集(40)	测试集(54)
AdaBoost	100%	91%
SVM	100%	90%
BP(神经网络)	100%	88%

从上表可以看出,对于训练样本,所有模型的分类准确率都是 100%;对于测试样本 AdaBoost 组合分类器分类准确率为 91%;SVM 模型的分类准确率为 90%;神经网络模型的分类准确率为 88%。组合分类器的分类准确率最高。

本文将 AdaBoost 算法引入上市公司信用风险评价中,建立了基于组合分类器的上市公司信用风险评价模型。实证结果表明,基于 AdaBoost 的组合分类器模型比单分类器模型具有更高的预测准确率。在信用评估领域,预测准确率即使只有微小的提升,也有可能给企业带来很大的收益,从这个角度看,本文的改进极具理论意义与应用价值。

主要参考文献

旷海兰.一种基于粗糙集理论的组合分类器构造方法.计算机工程与应用,2006;16