

doi: 10.12068/j.issn.1005-3026.2019.02.027

SVM 财务欺诈识别模型

曹德芳, 刘柏池

(东北大学 工商管理学院, 辽宁 沈阳 110169)

摘 要: 利用我国资本市场的面板数据,选取 2006—2015 年公布的财务报表欺诈公司作为样本公司,以 1:1 比例配比非财务欺诈公司,对 27 个指标(包括财务指标和非财务指标)进行分析,然后通过独立性检验对指标进行降维处理,最终保留 8 个建模指标. 分别利用网格搜索算法、遗传算法和粒子群算法进行支持向量机模型的参数寻优,基于上述不同算法建立了三个支持向量机财务欺诈识别模型. 最后,比较三个模型的运行效果,结果表明,通过粒子群算法寻找最优参数效果最好,据此建立的支持向量机模型可以很好地识别出财务欺诈公司样本.

关 键 词: 参数寻优;支持向量机;财务欺诈;识别模型;遗传算法;粒子群算法

中图分类号: F 830.91 **文献标志码:** A **文章编号:** 1005-3026(2019)02-0295-06

SVM Model for Financial Fraud Detection

CAO De-fang, LIU Bai-chi

(School of Business Administration, Northeastern University, Shenyang 110169, China. Corresponding author: CAO De-fang, E-mail: dfcao@mail.neu.edu.cn)

Abstract: Based on the panel data of China's capital market, the financial fraud companies from 2006 to 2015, together with the same number of non-fraud companies were selected as the research samples. Twenty-seven financial and non-financial indexes were analyzed, after which the dimension of the indexes was reduced through the test of independence and eight indexes were retained as the modeling parameters. The grid search algorithm, genetic algorithm and particle swarm optimization(PSO) were used respectively to optimize the parameters, and three support vector machine(SVM) models with the parameters optimized by the proposed methods were established respectively for financial fraud detection. The results showed that the SVM model with the parameters optimized by PSO has a higher detection rate than the other two models.

Key words: parameter optimization; support vector machine(SVM); financial fraud; detection model; genetic algorithm(GA); particle swarm optimization(PSO)

近年来,上市公司财务欺诈案件频发,给企业的长期可持续发展带来严重后果,也会对其雇员、投资者造成重大损失,威胁了我国资本市场的稳定;因此,研究有效的财务欺诈识别方法已经成为当务之急.

已有的财务欺诈识别方法研究从 Persons^[1]开始延续至今. 传统的财务欺诈识别方法有 Probit 回归、Logistic 回归、主成分分析回归、判别分析等. Cecchini 等^[2]利用支持向量机(SVM)分类方法建立了财务欺诈识别模型. 这个识别模型把 1991 年至 2000 年的早期数据作为训练集, 2001 年至 2003 年的后期数据作为测试集. 识别模型的训练集含有 107 家欺诈公司,按照大致 1:20 的比例选取了 2205 家非欺诈公司;测试集包含 25 家欺诈公司,按照大致 1:37 的比例选取了 982 家非欺诈公司. 实证结果表明:财务欺诈公司识别的准确度高达 80%,非欺诈识别精度高达 90.6%. 但是,文献[2]将早期数据作为训练集,后期数据作为测试集,识别率可能会受样本的时间因素影响,另外,两组数据的配比结构不同也会造成模型的不准确. 宋新平等^[3]从 2005 年 A 股制造业中选取 36 家财务欺诈公司作为欺诈样本,

选取另一些制造业非欺诈公司作为对照样本,另外选取 23 个财务指标,分别利用数据挖掘方法、多元判别分析、支持向量机、决策树,以及研究设计的集成分类方法构建了财务欺诈识别模型.结果表明这几种模型的识别效果都很好,其中,集成分类方法的识别效果最好.但是该研究的指标选取均为财务指标,没有考虑非财务指标对欺诈公司识别的敏感性;此外,这项研究的样本量过小,而基于支持向量机的人工智能模型要有一定数据量为基础,才能构造出足够稳定并且推广能力强的模型.

SVM 由 Vapnik^[4]于 1996 年提出,与常用的人工智能方法包括遗传算法、神经网络等方法相比,它具有充分的理论基础、优良的泛化能力、实用的非线性处理能力,以及强大的高维度数据处理能力.建立 SVM 模型时,参数选取对模型的分类能力有重要影响,许多学者对此进行了深入研究.文献[5]使用网格搜索算法选取 SVM 惩罚参数和径向基(RBF)核函数.由于网格搜索算法过程简单、容易理解,因此成为应用最广泛的算法,但是该算法得到的模型精度有限.为了解决 SVM 参数寻优问题,文献[6]用遗传算法(GA)进行 SVM 参数寻优,减少了训练模型所需的时间.由于 SVM 能够很好地解决大数据集的分类和回归问题,国内外的学者对 SVM 进行了持续的研究,SVM 方法日趋完善.除了传统的 SVM 模型以外,还出现了各种 SVM 综合模型,比如 Col - SVM^[7], TW - SVM^[8], NP - SVM^[9]等.文献[10]用人工蜂群算法对 SVM 的模型参数进行了优化.本文分别使用网格搜索算法、遗传算法(GA)、粒子群算法(PSO)进行 SVM 参数寻优,在传统 SVM 模型基础上,采用不同的参数寻优算法,对模型进行优化.

根据上述分析,本文提出如下的研究假设:基于 SVM 建立的财务欺诈识别模型可以提高财务欺诈的识别准确度,并且不同的参数选取方法对财务欺诈识别的准确度有所影响.

1 研究设计

1.1 样本选取及数据来源

本文的训练集和测试集中的数据都采用同时期的数据,排除了实验中时间因素的影响.同时,在样本选取过程中,找到了从 2006 年至 2015 年 98 家欺诈公司的样本数据,尽可能在数据量上达到人工智能模型构建的标准.本文选取我国沪深

两市 A 股上市公司非金融行业 2006—2015 年有财务欺诈行为的上市公司作为欺诈样本,具体要求如下:

1) 已经被证监会、上海证券交易所、深圳证券交易所披露,即在信息披露公告和市场禁入公告中被指出在 2006 年至 2015 年的年度会计报告中存在虚假信息披露、严重误导性陈述、重大错报和漏报现象.

2) 样本中剔除中期财务报告中存在财务欺诈现象但年度财务报告中没有欺诈现象的上市公司.

在选取对照样本时以 1:1 比例配对,具体配对样本的选取标准为:

1) 与财务欺诈公司同年份、同行业、资产规模相近的 A 股上市公司.

2) 同会计年度的审计意见为无保留意见的上市公司.

3) 非 ST,PT 的上市公司.

4) 没有被证监会、财政部、两证券交易所作出任何形式处罚的上市公司.

剔除数据不全的公司,最后有 98 家欺诈公司及 98 家对照公司数据,数据来源于国泰君安数据库、上海证券交易所、深圳证券交易所及证监会官方网站.

1.2 研究变量的初选

为了更加客观地衡量支持向量机的分类能力,本文在选取指标时分析国内外对财务欺诈识别指标的研究,综合考虑我国上市公司的特点,分别从盈利能力、偿债能力、经营能力、发展能力、现金流能力、股权结构、外部评价七个方面选取了共 27 个公司指标,全面衡量上市公司的情况.

1.2.1 财务指标

偿债能力体现了公司到期能够偿还债务的实力,往往负债率较高的公司,出现财务欺诈行为的可能性越高.本文从衡量长短期负债能力入手各选取了两个指标:流动比率、速动比率、资产负债率,以及有形资产负债率.

发展能力是指公司可持续经营的能力.本文选取的指标有:总资产增长率、净资产收益率增长率,以及净利润增长率.

经营能力也可称为营运能力,揭示了公司资金周转的情况,能够体现公司运用现有的各项资产赚取利润的能力.该能力与偿债能力和盈利能力有一定的勾稽关系.本文选取的指标有:总资产周转率、存货周转率、应收账款周转率、流动资产周转率.盈利能力体现了公司获取利润的能力,一

般来说,盈利能力越强,进行财务欺诈的可能性越低. 本文选取具有代表性的指标:资产报酬率、总资产净利率、净资产收益率,以及营业利润率.

现金流量指标在一定程度上也能体现公司进行财务欺诈的可能性. 公司想要对财务报表进行粉饰,通常会虚拟交易,增加公司利润,然而经营性现金流不会随之增加,选择净利润现金净含量、营业收入现金净含量及全部现金回收率这三项指标进行考察,如果指标异常,则公司很有可能通过虚增利润来进行财务欺诈.

1.2.2 非财务指标

随着财务欺诈识别模型研究的深入,国内外许多学者发现,公司的经营业绩不仅体现在财务指标上,非财务指标对经营业绩也会有所影响. 本文从公司所处的内部和外部环境入手,选取恰当的非财务指标.

公司的内部环境指标主要体现在公司的治理结构和股权结构,因此本文在股权结构方面选取国有股、流通股、法人股比例以及股权集中度等四项指标,而在治理结构方面分别选取董事、监事和股东三大会议召开的次数作为指标.

外部评价对于衡量上市公司情况非常重要,本文选取审计意见类型及是否被 ST 这两项指标,并对这两项指标进行量化:被 ST 为 1,不被 ST 为 0;标准无保留意见为 0,带强调事项段无保留意见为 1,无法表示意见为 2,保留意见为 3.

指标 $X_1 \sim X_4$ 代表偿债能力, $X_5 \sim X_7$ 代表成长能力, $X_8 \sim X_{11}$ 代表经营能力, $X_{12} \sim X_{15}$ 代表盈利能力, $X_{16} \sim X_{18}$ 代表现金流量, $X_{19} \sim X_{21}$ 代表内部控制, $X_{22} \sim X_{23}$ 代表外部评价, $X_{24} \sim X_{27}$ 代表股权结构. 指标含义如表 1 所示,

表 1 指标体系
Table 1 Index system

变量	指标名称	变量	指标名称
X_1	流动比率	X_{15}	营业利润率
X_2	速动比率	X_{16}	净利润现金净含量
X_3	资产负债率	X_{17}	营业收入现金含量
X_4	有形资产负债率	X_{18}	全部现金回收率
X_5	总资产增长率	X_{19}	董事会议次数
X_6	净资产收益率增长率	X_{20}	监事会议次数
X_7	净利润增长率	X_{21}	股东大会次数
X_8	应收账款周转率	X_{22}	审计意见类型
X_9	存货周转率	X_{23}	是否 ST
X_{10}	流动资产周转率	X_{24}	第一大股东持股
X_{11}	总资产周转率	X_{25}	国有股比例
X_{12}	资产报酬率	X_{26}	法人股比例
X_{13}	总资产净利率	X_{27}	流通股比例
X_{14}	净资产收益率		

1.3 研究变量的筛选

若选取的财务欺诈识别指标对欺诈公司和非欺诈公司的度量没有明显差异,则在之后建立识别模型时会严重影响识别准确度,因此将差异性不突出的指标予以剔除. 而且,虽然支持向量机 (SVM) 模型可以处理输入的高维变量,但是这种方式有两个弊端:第一,会增加建模计算量;第二,会影响模型的准确性. 因此,在构建模型之前要对变量进行差异分析,以便对数据降维.

首先利用 SPSS 对财务所有指标进行 T 对照样本独立检验,然后保留显著性小于 0.1 的指标,结果使变量降维到了 8 项,极大地降低了模型的复杂度.

保留下的指标中,财务指标分别为:流动比率 X_1 ,速动比率 X_2 ,流动资产周转率 X_{10} ,全部现金回收率 X_{18} . 非财务指标有:审计意见类型 X_{22} ,是否 ST X_{23} ,第一大股东持股比例 X_{24} ,国有股比例 X_{25} . 其中,审计意见类型 X_{22} 、国有股比例 X_{23} 、第一大股东持股比例 X_{24} 的显著性检验结果均小于 0.05,表明这三项指标对欺诈行为的敏感性非常高,同时也说明了非财务指标对财务欺诈行为的识别是有效的,这也与以往的研究结果相符合. 可见,将非财务指标引入财务欺诈识别指标体系是合理与可行的. T 检验结果如表 2 所示,

表 2 T 检验指标值
Table 2 T-Test of Indexes

变量	指标名称	T 值	显著性(双尾)
X_1	流动比率	2.322	0.022
X_2	速动比率	1.686	0.095
X_3	资产负债率	0.881	0.381
X_4	有形资产负债率	0.846	0.400
X_5	总资产增长率	0.036	0.972
X_6	净资产收益率增长率	0.518	0.605
X_7	净利润增长率	1.15	0.253
X_8	应收账款周转率	-1.06	0.292
X_9	存货周转率	-1.305	0.195
X_{10}	流动资产周转率	1.74	0.085
\vdots	\vdots	\vdots	\vdots
X_{18}	全部现金回收率	2.428	0.017
X_{19}	董事会议次数	-0.954	0.342
X_{20}	监事会议次数	-1.312	0.193
X_{21}	股东大会次数	0.222	0.825
X_{22}	审计意见类型	-6.654	0.000
X_{23}	是否 ST	-3.723	0.000
X_{24}	第一大股东持股	3.621	0.000
X_{25}	国有股比例	2.019	0.046
X_{26}	法人股比例	0.281	0.779
X_{27}	流通股比例	-0.602	0.548

1.4 数据归一化

因为建立支持向量机 (SVM) 模型是在 MATLAB 上运行,且使用的是 C++ 语言,因此,为了提高训练的效率,要在构建模型之前对初始数据归一化处理,将数据转化为 $[-1,1]$ 之间,缩小数据的波动范围,以便提高支持向量机精确度.记 v 是原始数据, v' 是归一化后的数据, \max, \min 分别代数据中每个指标的上下界;具体表达式为

$$v' = \frac{v - \min}{\max - \min} \quad (1)$$

2 模型构建及检验

2.1 SVM 模型的构建流程

构建 SVM 模型需要选取核函数,核函数种类主要包括径向基核函数、多项式核函数等,本次模型选用径向基核函数,径向基核函数最主要的优点是参数少、适用性强.确定核函数后,在构造模型时还需要选取核函数的参数.应用三种方法选择参数,并比较这三种方法选择参数的准确度和代入检测样本后得到识别率的准确度.选用三种方法进行参数选择,是因为在已有的利用支持向量机进行财务欺诈识别的研究中,基本上都是直接利用简单模型,而对不同的参数选取方法没有做过详细的比较.实际上,针对不同的数据类型和来源,参数选取的方法会直接影响模型的准确率和适用推广能力;本文的研究旨在针对这部分缺失进行补充和探讨,这正是本文的学术贡献所在.

具体流程为,将 96 对样本分成两组,前 45 对为训练样本,其余 51 对为测试样本.在 MATLAB 及 VS2010 平台上编译.分别运用网格搜索算法、遗传算法、粒子群算法通过训练样本来得到径向基核函数的最优参数,应用序列最小优化算法解决计算中的二次优化问题.得到参数后,分别建立三种模型,用测试样本对三种模型进行检验,得到财务欺诈样本的识别准确率.

2.2 模型的建立与检验

本文在台湾大学林智仁教授开发出的 Libsvm 工具包的基础上采用 C++ 语言实现支持向量机模型的仿真研究.选取径向基核函数,惩罚因子 C 与径向基核函数的参数 g 分别运用网格搜索算法、遗传算法 (GA)、粒子群算法 (PSO) 确定.

2.2.1 网格搜索算法及交叉验证思想

网格搜索算法的基本原理是以惩罚因子 C

和核函数参数 g 为坐标,在两个参数的取值范围内按一定单位划分出网格,同时遍历网格内所有节点,在每个节点取值,得到参数 C 和 g 后,再使用 K 阶 (本文采用 10 阶) 交叉验证的方法得出该组取值 C 和 g 时训练集的分类准确率.

交叉验证 (cross validation, CV) 的功能是在一定范围内合理地选出支持向量机模型最优核函数参数 g 和最优惩罚因子 C ,并且在这一过程中有效避免过度拟合的发生.该实验中采用了 8 折交叉验证,将数据集分为 8 份,依次将其中 7 份作为训练数据,剩下的 1 份作为测试数据,然后试验.最终选出使得训练集分类结果准确率高的那组 C 和 g 作为生成分类模型的最佳参数.本次实验中,参数 C 的取值范围设为 $[2^{-10}, 2^{10}]$,参数 g 的取值范围设为 $[2^{-10}, 2^{10}]$,运行步长为 0.5,使用网格搜索算法,结果如图 1 所示.可知最佳参数 $C = 147.033\ 4, g = 9.189\ 6$,交叉验证精度为 77.272 7%.

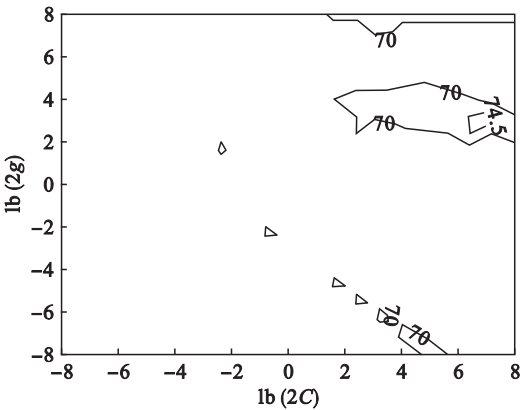


图 1 网格搜索算法参数结果
Fig. 1 Result of grid search method

2.2.2 遗传算法

遗传算法 (genetic algorithm, GA) 通过计算可得到全局最优解,且收敛速度快,是一种应用普遍的人工智能方法.遗传算法主要原理是通过对生物进化机制的模仿实现对模型的改进.遗传算法包括选择、交叉或基因重组、变异这三个基本步骤,支持向量机在创立之初的检测过程中,参数的选择都使用默认参数,针对性不强,识别准确率不高.遗传算法的并行处理和全局搜索能力使得对算法的参数进行优化运行时过程很短,并能寻找出全局最优解.在 MATLAB 上用遗传算法对支持向量机模型进行参数优化选取,最终结果如图 2 所示.结果显示:终止代数 为 199,种群数量为 20;最佳参数 $C = 12.689\ 6, g = 208.120\ 5$,交叉验证精度为 71.818 2%.

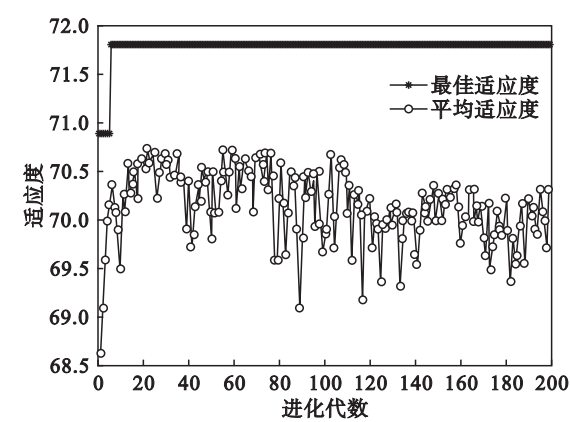


图 2 GA 算法适应度曲线
Fig. 2 Fitness curve of GA

2. 2. 3 粒子群算法

粒子群 (particle swarm optimization, PSO) 算法是一种新型群体智能算法. 相比于遗传算法, 该算法概念易懂, 计算量小. 粒子群算法的适应度值由被优化函数决定, 该值是评价粒子优劣的标准. 适应度值 w 越大, 算法越偏向全局搜索; 适应值 w 越小, 算法越偏向局部搜索. 在迭代过程中, 适应度值 w 线性递减, 兼顾全局搜索和局部搜索两方面, 达到参数寻优的目的. 图 3 为粒子群算法的适应度曲线, 其中 $c_1 = 1.5, c_2 = 1.7$, 终止代数为 200, 种群数量为 20; 最佳参数 $C = 61.1444, g = 12.6998$, 交叉验证精度为 76.3636%.

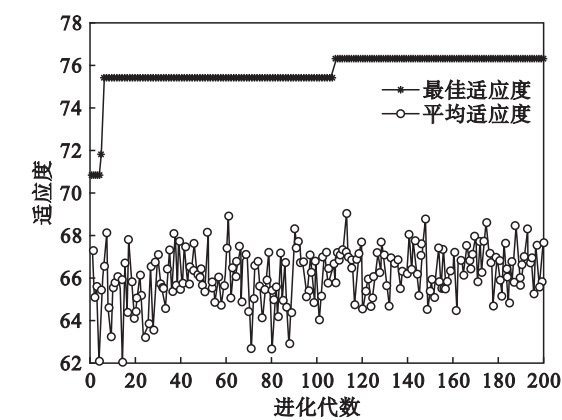


图 3 粒子群算法适应度曲线
Fig. 3 Fitness curve of PSO

使用三种方法寻找到的参数建立支持向量机模型, 将训练得到的模型代入 SVM - train 函数, 对测试样本进行识别, 识别结果如表 3 所示.

表 3 的数据表明, 三种方法的训练精度水平接近, 从训练时间上来看, 网格搜索算法速度最快, 遗传算法和粒子群算法速度都比较慢; 而在模型的识别效果方面, 网格搜索算法和粒子群算法都较精确, 遗传算法的测试精度和训练精度相差

较大, 说明遗传算法的结果不够稳定, 推广能力不如网格搜索算法和粒子群算法. 粒子群算法的训练精度和测试精度相差最小, 说明粒子群算法稳定性好, 推广能力强. 综上, 在利用支持向量机 (SVM) 建立财务欺诈模型上, 粒子群算法优于另外两种算法.

表 3 SVM 模型的识别准确率 Table 3 Accuracy of SVM model			
性能指标	网格算法	遗传算法	粒子群算法
训练精度/%	77.2727	71.8182	76.3636
测试精度/%	71.8182	65.1163	74.4186
运行时间/s	13.127	25.042	27.978

3 结 语

本文以上市公司的面板数据为研究数据, 选取了上市公司欺诈公司样本和配对公司样本, 选择合适的指标, 通过 T 独立样本检验降维, 然后将得到的指标数据归一化后建立 SVM 模型. 通过网格搜索算法、遗传算法和粒子群算法分别得到最优参数并建立 SVM 模型.

实证结果表明, 三种算法中, 粒子群算法有较好的识别准确率, 同时, 粒子群算法的泛化能力也很强. 综上, 粒子群算法在对财务欺诈识别上具有很实用的价值. 但是粒子群算法也存在不足, 这种方法在运行时明显需要更长的时间.

在下一步研究中, 可以在提高算法运行的效率上进一步探索, 提高模型的实用性. 此外, 对 SVM 的研究可以从基础模型衍生出更多模型, 比如参数选择上交叉利用蜂群算法等新算法; 最后, 是对 SVM 基本模型的改进, 比如非平行平面的支持向量机、Fisher 改进 SVM、最小二乘 SVM 等. 对 SVM 模型的综合改进是现行研究的主要趋势.

参考文献:

[1] Persons O S. Using financial statement data to identify factors associated with fraudulent financial reporting [J]. *Applied Business Research*, 1995, 11(3): 38-46.

[2] Cecchini M, Aytug H, Koehler G J, et al. Detecting management fraud in public companies [J]. *Management Science*, 2010, 56(7): 1146-1160.

[3] 宋新平, 丁永生, 张革夫. 集成分类法在财务欺诈风险识别中的应用 [J]. *计算机工程与应用*, 2008, 44(34): 226-230.

(Song Xin-ping, Ding Yong-sheng, Zhang Ge-fu. Application of integrated classification method in identifying risk of fraudulent financial report [J]. *Computer Engineering and Applications*, 2008, 44(34): 226-230.)