量化策略

2017 年 12 月 15 日

数据挖掘在上市公司财务造假识别中的应用

量化策略

相关研究

《上市公司财务造假预测模型研究》，2017 年 10 月

证券分析师
曹春晓 A0230516080002
caocx@swsresearch.com

研究支持
文 雨 A0230117100001
wenyu@swsresearch.com

联系人
文 雨
（8621）23297818×7397
wenyu@swsresearch.com

申万宏源研究微信服务号

投资提示：

- 上市公司财务造假犹如地雷，1999 年至 2002 年间，美国市场爆发出大量的财务造假案件，对投资者造成了不可估量的损失。2011 年，中概股造假案件高发，数家中概股因财务舞弊在美国两大证券交易所停牌或被勒令退市。国内市场近年来也频繁爆发惊天骗局，层出不穷的上市公司造假案件给投资者带来了巨大损失，也给资本市场的健康发展带来负面影响。
- 我们选取 2002 年之后，A 股市场被中国证监会、沪深两市交易所公开处罚的数据，依据财务造假常见的动机和手段，结合国内外学者的研究，构建了多个财务指标和非财务指标，并分别采用数据挖掘中的神经网络、SVM 支持向量机、决策树（C&RT、QUEST、CHAID、C5.0），对上市公司年报造假判别问题进行研究。
- 实证研究结果表明，由于样本数据的分类不平衡问题，决策树比神经网络、SVM 支持向量机更适用于上市公司年报造假判别的研究，而决策树中的 CHAID 算法是四个算法中综合表现较优的算法。在全样本中，其对上市公司年报是否造假判别的准确率为 93.16%，召回率为 59.41%，判断的精确度为 17.78%，F 值为 27.37%。
- 此外，在判断上市公司年报是否造假的时候，需要重点关注的几项指标分别为：审计师意见、前一年是否亏损、其他应收款占流动资产比例、销售毛利率、预付款项占流动资产比例。审计师给出负面意见、前一年财务亏损、相对于同行业而言，比较高的其他应收款占流动资产比例、销售毛利率、预付款项占流动资产比例等，都是造假可能性高的上市公司年报所具有的特征。

目录

1、研究概述

上市公司财务造假犹如地雷，1999 年至 2002 年间，美国市场爆发出大量的财务造假案件，对投资者造成了不可估量的损失。2011 年，中概股造假案件高发，数家中概股因财务舞弊在美国两大证券交易所停牌或被勒令退市。近年来国内市场也频繁爆发惊天骗局，层出不穷的上市公司造假案件给投资者带来了巨大损失，也给资本市场的健康发展带来负面影响。

回顾国内外造假案例，常见的财务造假手段包括：（1）虚增交易：通过伪造销售合同、销售发票等原始单据，编制虚假代销清单，虚构交易，并形成虚假收入和利润。（2）虚增资产：虚增资产主要的表现形式就是对一些已经没有利用价值的项目不予注销。（3）提前确认收入，虚增收入：未销售出去的商品或劳务，提前确认销售收入。（4）利用过渡性科目，调节利润：调整跨期费用，将一些已实际发生的费用作为长期待摊费用、待处理财产损失、其他应收款等项目入账，而不按照相关准则要求计入当期损益。（5）隐瞒或不及时披露重大事项：母子公司之间的关联方交易。母子公司之间往往利用不公允的市场价格，高买低卖，以此来达到操纵利润的目的。资产、债务重组上市公司常常利用其母子公司进行资产或债务重组。将自己的不良资产转让给关联公司，母子公司将优良资产输送给上市公司，从而达到输血的目的。

上市公司财务造假案件曝光后，大多数公司难逃厄运。在发展较为成熟的海外市场，投资者对于财务造假案件容忍度更低，众多昔日的明星公司，因财务造假曝光而退市或破产。对于大量购买了造假公司股票的投资者而言，其损失不言而喻。因此，如何识别那些可能进行财务造假的公司，是分析师、投资者和监管机构共同关心的重要问题。

国内外学者已经对如何判别上市公司年报等财务报告是否造假问题展开了相关研究。比如，Loebbecke 和 Willingham（1988）从管理层的角度出发进行分析研究，其研究结果表明，管理层是否存在造假的动机是对公司进行财务违规识别时需要注意的问题。Persons（1995）采用回归模型对财务报告违规行为特征进行识别，其研究结果显示，上市公司财务报告违规与资本周转率负相关，与财务杠杆和流动资产比率正相关。随着计算机技术的发展，数据挖掘相关技术也被一些学者运用到财务造假的研究当中。比如 Fanning 和 Cogger(1998)利用神经网络模型研究财务报告违规识别问题，他们从管理者出发，使用具有较高识别能力的指标，既包含了财务指标，也包含了非财务指标。他们认为，财务报告利益相关者可以根据该模型对是否发生财务报告违规进行初步的判断，根据初步判断的结果对所有上市公司分情况增加或减少现场检查。Kirkos 和 Spathis（2007）等对数据挖掘技术中的决策树模型、神经网络模型等进行了对比分析，以探究各模型的识别效果优劣。文章的实证分析结果表明，这些数据挖掘技术均取得了不错的分类效果。

本文将分别采用神经网络、SVM 支持向量机和决策树等数据挖掘方法来对上市公司年报造假的识别展开研究，并比较分析这些数据挖掘技术的相关表现，从而选择出最适用于识别上市公司年报造假的数据挖掘方法。

另外，采用数据挖掘方法进行研究的过程中，比较容易遇到的一个问题是单个模型可能具有数据依赖性，如何才能获得一个具有准确、稳定判别效果的模型是一个值得思考的问题。而这个问题一定程度上可以通过采用 Bagging 或者是 Boosting 的方法来解决，其中 Bagging 是 Bootstrap Aggregating 的一种方式，是一种根据均匀概率分布从数据集中重复抽样（有放回的）的技术，每个自助样本集的大小都和原数据集相同。由于抽样过程是有放回的，因此一些样本可能在某一个训练数据集中出现多次，而其它一些却可能被忽略。对于最后的结果判定，如果是分类问题采用投票方式，对回归问题则采用简单平均方法。

Boosting 主要是 AdaBoost (Adaptive Boosting)，初始化时对每一个训练集赋予相等的权重，然后用该算法对训练集训练 t 轮。每次训练后，对训练失败的数据赋以较大的权重，即让算法在后续的学习中重点对比较难的训练数据进行学习。从而得到一个预测函数序列 h_1，…，h_m，其中 h_i 也有一定的权重，预测效果好的预测函数权重较大，反之较小。最终的预测函数 H 对分类问题采用有权重的投票方式，对回归问题采用加权平均的方式对新数据进行判别。

Bagging 和 Boosting 都可以有效地提高分类的准确性。在大部分数据中，Boosting 的准确性比 Bagging 高；但在有些数据集中，Boosting 会引起过度拟合。

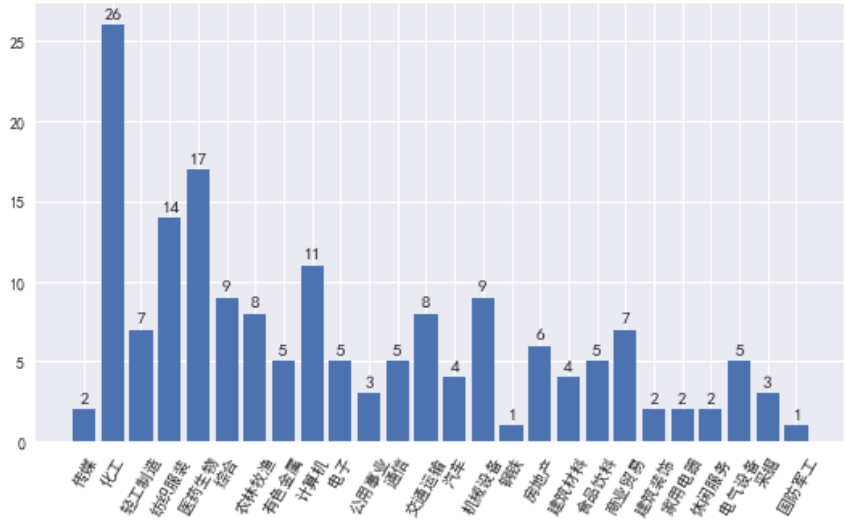
为了保证模型的稳定性，避免过度拟合，本文在后续研究中分别在神经网络模型、决策树中的 C&RT、QUEST、CHAID 采用 Bagging 方法来增强模型的稳定性。

2、数据说明和处理

2.1 数据说明

我们选取的造假样本为 2002 年之后，A 股市场被中国证监会、沪深两市交易所公开确定为财务造假的上市公司。剔除 IPO 之前以及新股上市当年造假的数据，对于同一家上市公司连续多个年度造假的数据，仅保留其首次造假年份作为研究数据，再剔除 2016 年年报数据之后，共 171 条记录。部分上市公司还存在季报或临时报告中财务数据造假的情况，但此类报告中的部分财务指标与年报不统一，故在此不纳入造假样本之列。

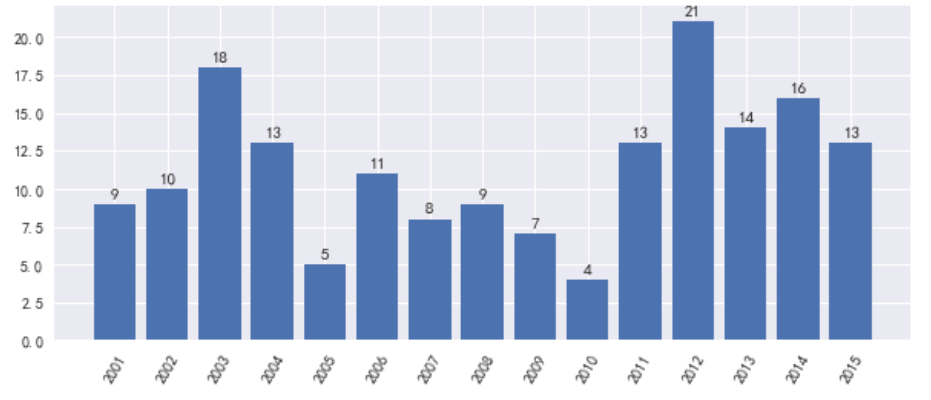
图 1 造假样本的行业分布图



资料来源：申万宏源研究，Wind

从上市公司年报造假样本的行业分布特征来看，化工、医药生物、纺织服装等行业涉及公司数量较多；从造假样本年份分布情况来看，2011-2015 年，是案件高发期。

图 2 造假样本的造假年度分布图



资料来源：申万宏源研究，Wind

我们采用控制样本匹配法，选取造假样本当年同行业所在的上市公司（剔除被证监会、交易所处罚的公司）作为对照样本。由造假样本和对照样本组成的全样本总共有 7839 条记录。

2.2 数据处理

根据国内外学者对于上市公司财务造假现象的研究结论，尽管上市公司对财报数据进行了粉饰，但是一些财务指标的异常还是为财务造假的侦查提供了线索。我们结合公司参与年报财务造假的动机、常见的手段，重点以财务指标为研究对象。下表所列为我们初步筛选的特征指标，由于不同公司财务数据差别较大，我们尽可能地使用比率数据进行研究。此外，根据上篇报告结论，审计师意见、前一年度是否亏损、前五大股东占比等指标也相对有效，也纳入特征指标列表中。

表 1：特征指标列表

名称	代号	名称	代号
资产负债率	Asset_lia_ratio	净利润同比增长率	Npro_gro_rate
流动比率	Curr_ratio	主营业务收入同比增长率	Rev_gro_rate
速动比率	Quick_ratio	营业利润同比增长率	Opro_gro_rate
主营业务毛利率	Mainb_grop_rate	经营活动产生的净流量增长率	Ntra_gro_rate
营业收入净利润率	Rev_netp_rate	应收账款占流动资产比例	Rec_to_Cur
主营业务利润占比	Mainb_pro_ratio	预付款项占流动资产比例	Pre_to_Cur
销售净利率	Sale_pro_ratio	其他应收款占流动资产比例	OtherRec_to_Cur
销售毛利率	Sale_grop_ratio	净资产收益率	ROE
货币资金占流动资产的比率	Monetary_cur_ratio	总资产报酬率	ROA
存货占流动资产的比率	Stock_cur_ratio	应收账款周转率	Rec_tur_rate
流动资产比率	Curr_ass_ratio	存货周转率	Sto_tur_rate
无形资产比率	Invisible_ass_ratio	总资产周转率	Tola_tur_rate
营业总收入/净利润	Rev_to_Npro	造假前 1 年是否亏损	Pre_np
营业外收入/净利润	Nonb_to_Npro	前五大股东股权集中度	Hold_pct
应收账款/营业收入	Rec_to_Rev	审计信息	Opinion
营业收入现金净含量	Cash_to_Rev	应计项	TATA
营业利润现金净含量	Cash_to_Pro		

资料来源：申万宏源研究

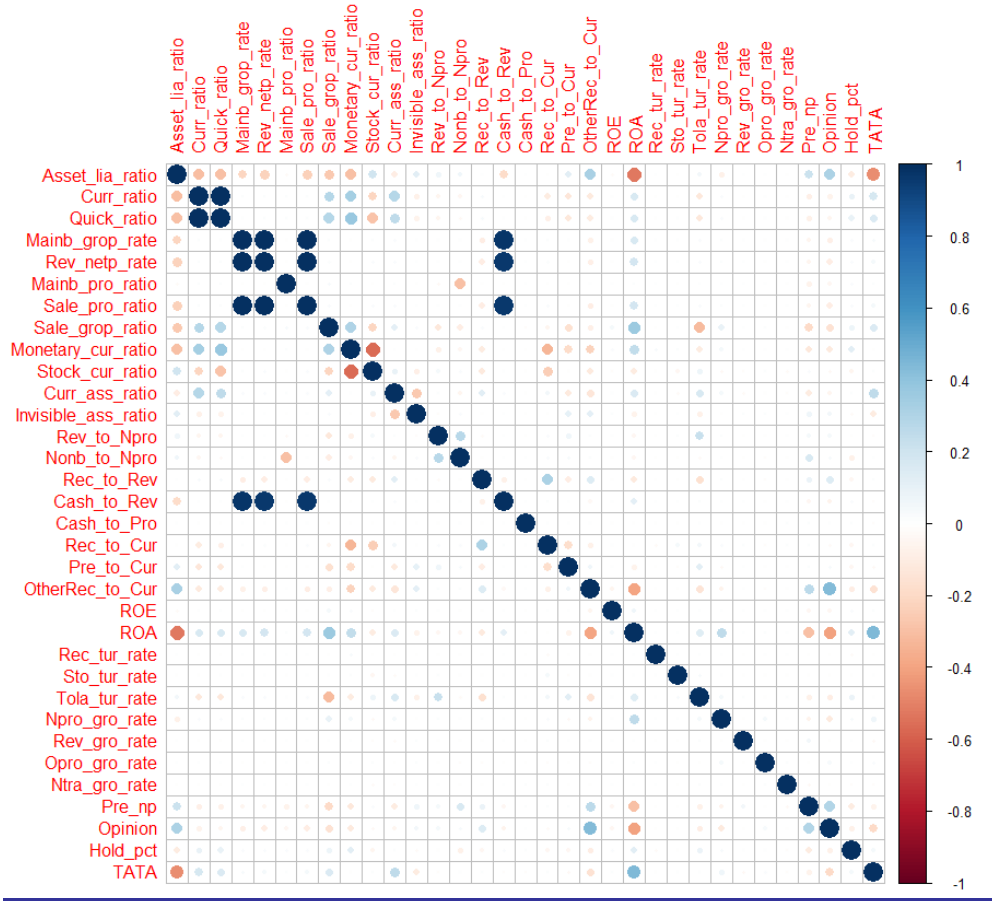
数据不平衡问题是使用数据挖掘方法时较为常见的问题，例如本文中我们所选的造假样本和控制样本的比例约为 1:46，即控制样本（非造假样本）的数量大于造假样本的数量。处理数据不平衡问题的方法很多，比如对数据集进行重采样、使用 SMOTE 方法来构造人工数据样本、设置不同的误分类损失或者是把数量相对过小的类别作为异常点，从而转化为异常点检测问题等。

由于在实际投资中，把一个真正造假的上市公司年报判断为非造假年报所导致的成本远大于把一个非造假的上市公司年报判断为造假年报所带来的成本。因

此，在处理样本类别不平衡的问题时，本文采用的是设置不同的误分类损失方法，即假设年报造假样本标为 1，非造假样本标为 0，则把 1 分类样本预测为 0 分类的成本是把 0 分类样本预测为 1 分类的 50 倍。

此外，特征指标之间往往存在一定的相关关系，如下图所示，资产负债率（Asset_lia_ratio）与流动比率（Curr_ratio）、速动比率（Curr_ratio）、总资产报酬率（ROA）等之间的负相关性较高；而流动比率与速动比率的正相关性非常高；主营业务毛利率（Mainb_grop_rate）与营业收入净利润率（Rev_netp_rate）、销售净利率（Sale_pro_ratio）等指标之间的正相关非常高，这些结论与常识相符。根据某个变量与其他变量之间的相关性程度大小，以及相关变量个数多少为原则，剔除一些相关程度大并且相关变量多的变量，即资产负债率（Asset_lia_ratio）、流动比率（Curr_ratio）、营业收入净利润率（Rev_netp_rate）、销售净利率（Sale_pro_ratio）、货币资金占流动资产的比率（Monetary_cur_ratio）、营业外收入/净利润（Nonb_to_Npro）、营业收入现金净含量（Cash_to_Rev）、总资产报酬率（ROA）等。

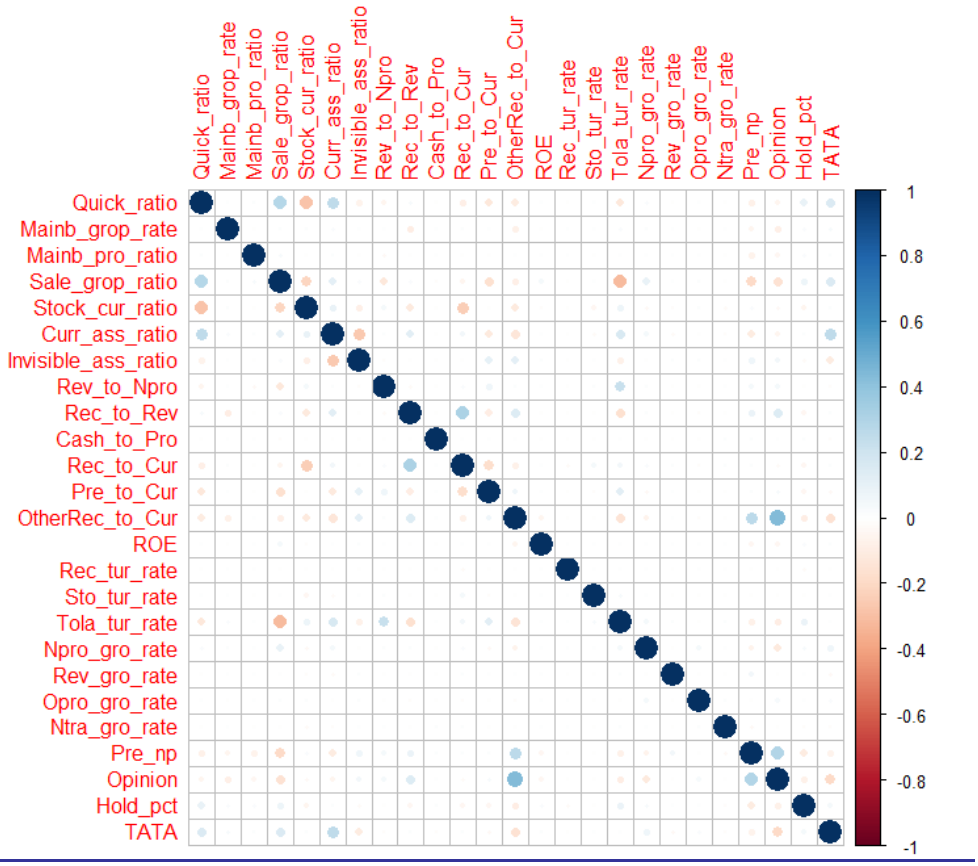
图 3：特征指标间相关关系图



资料来源：申万宏源研究

剔除上述指标之后，剩下的指标之间相关情况如图 4 所示：

图 4 选定指标之间相关情况图



资料来源：申万宏源研究

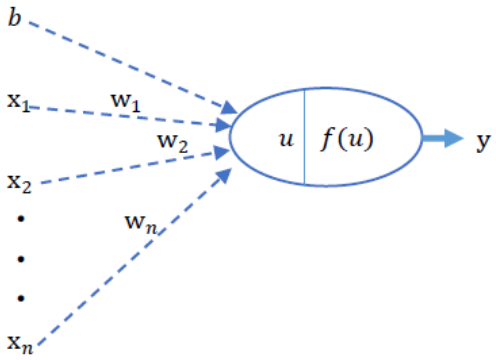
接下来，本文以随机选取的方式把全样本按照 7:3 的比例分为训练集和测试集，并分别用神经网络模型、SVM 模型和四种不同算法的决策树模型对上市公司年报财务造假识别问题进行研究。

3、多层神经网络

3.1 方法简介

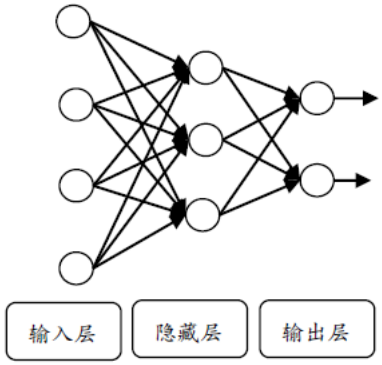
神经网络（Neural Networks）是一种模仿生物神经系统的机器学习算法，与生物神经系统相似，人工神经网络也是由若干个神经元构成。

图 5：神经元



资料来源：申万宏源研究

图 6：神经网络结构图



资料来源：申万宏源研究

如图 5 所示，x_1、x_2、…x_n为该神经元的输入，y为该神经元的输出。显然，不同的输入对神经元的作用是不同的，因此用权值w_1、w_2、…w_n来表示这种影响程度的不同。神经元内部包括两个部分，第一个部分是对输入的加权求和，第二个部分是对求和的结果进行“激活”，得到输出。加权求和的公式为：

$$u = \sum_{i=1}^{n} (w_i x_i) + b$$

式中，b 为偏移量，该偏移量也可以定义为输入恒为 1 的权值w_0，即权值也包括偏移量，因此上式可以改写为：

$$u = \sum_{i=0}^{n} (w_i x_i)$$

激活的公式为：

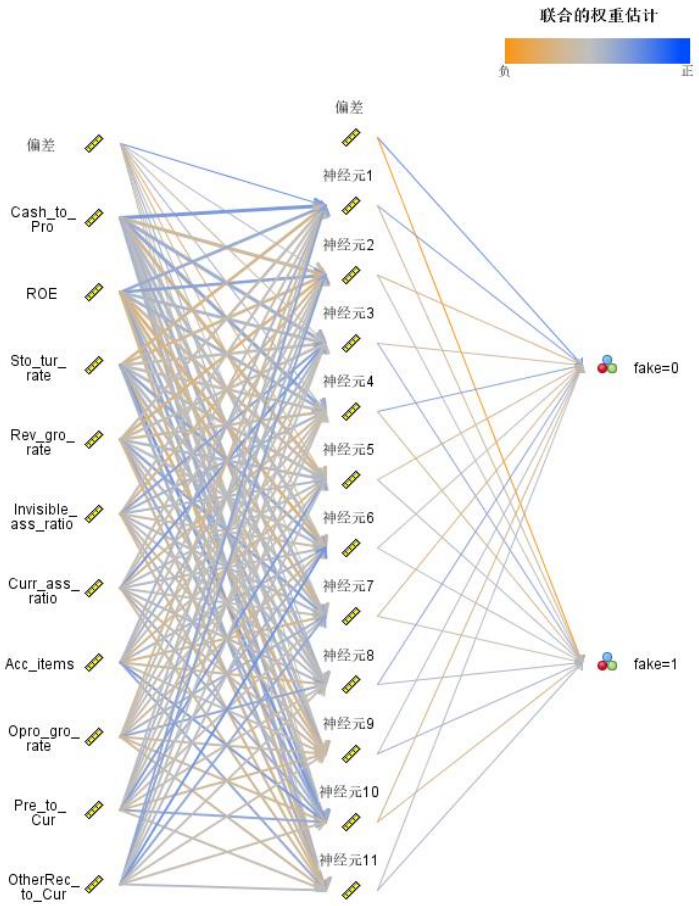
$$y = f(u)$$

其中，$f(\cdot)$称为激活函数，其可以有多种形式，比如线性函数、阈值函数、双曲正切函数等。而前馈神经网络是神经网络的一种，也是最常用的一种神经网络。它包括一个输入层，一个输出层和若干个隐含层，因此具有该种拓扑结构的神经网络又称为多层感知器（MLP）。如图 6 所示，该 MLP 包括一个输入层，一个输出层和一个隐藏层，其中某一层的神经元只能通过一个方向连接到下一层的神经元。MLP 是感知器的推广，克服了感知器不能对线性不可分数据进行识别的弱点。

3.2 模型结果

使用多层感知器（MLP）算法对上市公司年报造假进行识别分析，当构建单个神经网络时，计算所得的神经网络其中一部分结构图如下图所示：

图 7：神经网络部分结构图



资料来源：申万宏源研究

由上图可知，经过模型构建之后，一共有一个输入层、一个隐藏层和一个输出层。隐藏层包括 11 个神经元和一个偏差项，每个神经元与输入层的各个变量之间的权重系数如图中连接线的粗细和颜色所示。越粗的线代表权重系数绝对值越大，蓝色代表权重系数为正值，橙色代表权重系数为负值。

为了获得一个比较稳定的模型，令 Bagging 次数为 10，即构建 10 个神经网络模型，对样本的最后判断结果由 10 个模型投票表决的结果产生。计算所得的整体模型预测准确性结果如表 2 所示：

表 2：神经网络模型预测准确性

分区	训练集	百分比率	测试集	百分比率
正确	5,329	97.92%	2,345	97.83%
错误	113	2.08%	52	2.17%
总计	5,442		2,397	

资料来源：申万宏源研究

由表 2 可知，神经网络模型在训练集和测试集里的准确率分别为 97.92%、97.83%。虽然模型获得了高的准确率，但是准确率只是判断模型有效性的其中一个评价指标。另外还需要结合混淆矩阵来进行综合判断。该模型的混淆矩阵如下表所示：

表 3：神经网络模型的混淆矩阵

训练集	0	1	测试集	0	1
0	5,324	0	0	2,345	0
1	113	5	1	52	0

资料来源：申万宏源研究，注：横坐标“0”、“1”分别表示预测结果为“非造假样本”、“造假样本”，纵坐标的“0”、“1”分别表示实际为“非造假样本”、“造假样本”。

由上表可知，MLP 神经网络模型虽然能够非常准确地判断出非造假样本。但是，此神经网络模型在训练集和测试集和全样本中的准确率、精度、召回率和 F 值如下表所示：

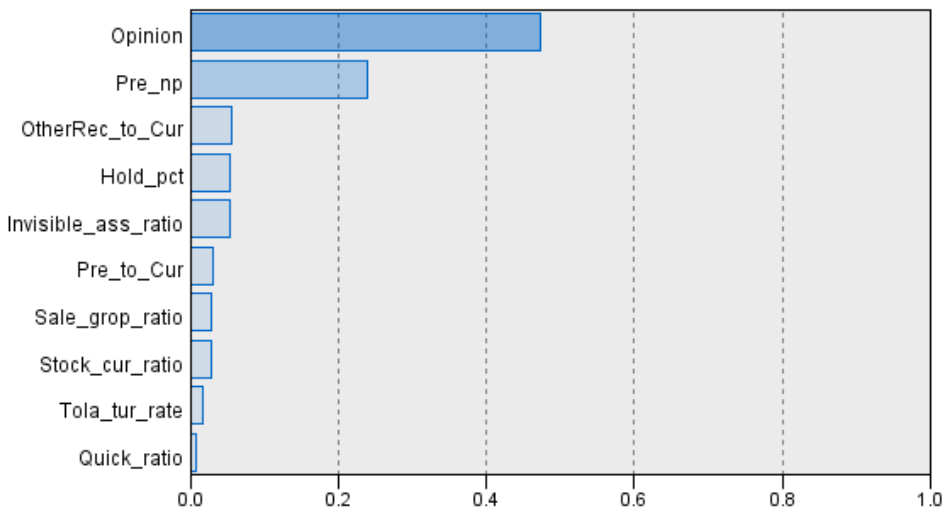
表 4：神经网络模型判别效果的评价指标结果

	训练集	测试集	全样本
准确率	97.92%	97.83%	97.90%
精确度	100.00%	无	100.00%
召回率	4.24%	0.00%	2.94%
F 值	8.13%	0.00%	5.71%

资料来源：申万宏源研究

其中，**准确率为预测对的样本占全样本比率；精确度是预测为造假样本中，真正是造假样本的比率；召回率时预测对的造假样本数量占所有造假样本数量的比率；F 值为精确度和召回率的调和均值。**由表 4 可知，虽然神经网络模型的准确率很高，但是由于样本数据分类的不平衡问题，实际上有很大比例的造假样本没有被识别出来。另外，各个预测变量的重要情况如下图所示：

图 8 神经网络中预测变量重要性



资料来源：申万宏源研究

由图 8 可知，重要性排名在前四的变量分别为 Opinion（审计信息）、Pre_np（造假前 1 年是否亏损）、OtherRec_to_Cur（其他应收款占流动资产比例）、Hold_pct（前五大股东股权集中度）。由此可知，当审计师给出负面意见时，其年报财务造假的可能性非常高。另外，由于中国股市 ST 制度的存在，当该上市公司前一年的财务出现亏损状况时，当年进行财务造假的动机也很非常高。而财务造假里面的常用手段就是通过关联交易或者其他方式虚增收入，这些往往在其他应收账款有所体现。最后，年报财务造假的上市公司大股东持股比例不高，以较低比例获得上市公司控制权。因为造假本质上就是大股东掠夺小股东财富的行为。因此，只有小股东占比越大，供大股东获利的空间才越大。另外，这也说明了把非财务指标和财务指标相结合来进行研究，更能准确地判断上市公司年报是否造假。

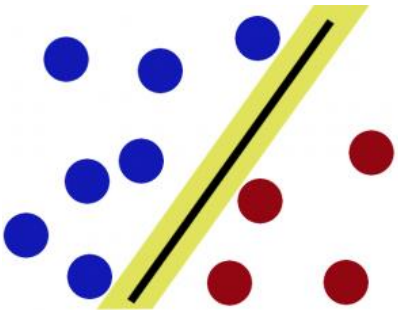
4、支持向量机 SVM

4.1 方法简介

支持向量机(Support Vector Machine，SVM)是 Corinna Cortes 和 Vapnik 等人于 1995 年首先提出的，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。支持向量机属于有监督学习模型，可以分析数据，识别模式，用于分类和回归分析。

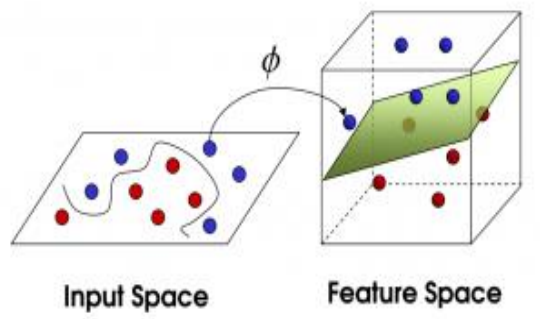
SVM 的分类算法：给定一组训练样本，每个标记为属于两类，SVM 训练算法建立了一个模型，分配新的实例为一类或其他类，使其成为非概率二元线性分类。例如，如图 9 所示，SVM 算法是求得一个最优的切线，使得其与两端最近的蓝色和红色点之间的距离尽可能大。但是，如图 10 所示，在有些情况下不能通过在一个二维平面上加入一条切线的方式去分开两类样本时，可以把二维平面映射到更高维的空间里，去寻找一个最优分类的切面。

图 9 SVM 线性分类例子



资料来源：申万宏源研究

图 10 SVM 非线性分类例子



资料来源：http://bytesizebio.net

从数学方程式上解释 SVM 支持向量机，则假设数据点 x 为 n 维向量，$w=[w_1,w_2,\cdots,w_n]$，b为常数，SVM 线性分类器的学习目标就是从 n 维数据空间中找到一个能够把数据进行分类的超平面，其方程表达式如下：

$$w^T x+b=0$$

其中$w^T x+b<0$为其中一个分类，而$w^T x+b>0$为另一个分类。当新的数据x_{new}进来时，计算$w^T x_{new}+b$的值，并根据该值的正负情况来划分到相关的分类中。因此，$f(x)=w^T x+b$即为线性分类函数。

另外，如前面例子所述，对于一些线性不可分的情况，SVM 的处理方法是选择一个核函数来将数据映射到更高维空间中进行划分。在 SVM 理论中，采用不同的核函数将导致不同的 SVM 算法。在本章 SVM 实证研究中，核函数采用的是使用最广的径向基核（RBF，Radial Basis Function）核函数。无论是小样本还是大样本，高维还是低维等情况，RBF 核函数均适用。

4.2 模型结果

SVM 模型所得结果的准确率如下表所示：

表 5：SVM 模型准确率

分区	训练集	百分比率	测试集	百分比率
正确	5,322	97.79%	2,340	97.62%
错误	120	2.21%	57	2.38%
总计	5442		2,397	

资料来源：申万宏源研究

由上表可知，SVM 模型在训练集和测试集中所得的准确率分别为 98.12%和 97.33%。与 MLP 神经网络模型的准确率相近。另外，混淆矩阵如下表所示：

表 6：SVM 模型的混淆矩阵

训练集	0	1	测试集	0	1
0	5,315	9	0	2,339	6
1	111	7	1	51	1

资料来源：申万宏源研究，注：横坐标“0”、“1”分别表示预测结果为“非造假样本”、“造假样本”，纵坐标的“0”、“1”分别表示实际为“非造假样本”、“造假样本”。

由该混淆矩阵计算得 SVM 模型判别效果的相关评价指标如下：

表 5：SVM 模型判别效果的评价指标结果

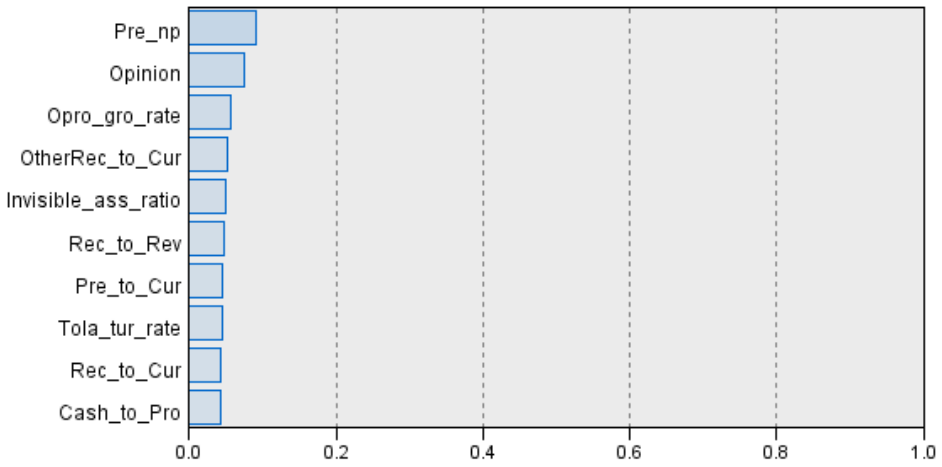
	训练集	测试集	全样本
准确率	97.79%	97.62%	97.74%
精确度	43.75%	14.29%	34.78%
召回率	5.93%	1.92%	4.71%
F 值	10.45%	3.39%	8.29%

资料来源：申万宏源研究

根据表 7 中的结果可知，由于样本数据的分类不平衡问题，SVM 模型在分类过程中，同样地对非造假样本具有很高的准确判断性，但是比较难以识别造假样本。

SVM 模型中变量重要性情况如下图所示：

图 11 SVM 模型中预测变量重要性



资料来源：申万宏源研究

由上图可知，重要性排名在前四的变量分别为 Pre_np（造假前 1 年是否亏损）、Opinion（审计信息）、Opro_gro_rate（营业利润同比增长率）、OtherRec_to_Cur（其他应收款占流动资产比例）。因此，这四个变量中有三个变量也在神经网络模型重要性排名前四的变量中，这两个模型所得的预测变量重要性相似。

5、决策树

5.1 方法简介

决策树是在进行数据挖掘时经常使用的分类和预测方法，该方法的基本原理为：通过算法中所规定的分类条件对于整体数据进行分类，产生一个决策节点，并持续依照算法规则分类，直到数据无法再分类为止。即，通过对训练样本的学习，建立分类规则；依据分类规则，实现对新样本的分类；属于有监督式的学习方法，有两类变量：目标变量，输入变量。

决策树算法根据其算法原理以及所适用分析数据类型的不同延伸出多种决策树算法。例如 C&RT、QUEST、CHAID 和 C5.0 等，这些算法的相关简介如下：

（1） C&RT 算法

又称为 CART，构建决策树的原理是使用 Gini 指数作为判定决策树是否仍须进行分支的依据，并建立二元的决策树，此算法不管是目标变量还是输入变量均适用连续或者分类类型。

（2） QUEST 算法

此算法是利用统计方法分割数据，即基于统计结果判定决策树是否仍需进行分支，以建立二元的决策树。QUEST 在变量的数据类型限制上，不适用于目标变量为连续类型，但输入变量则适用于连续和分类类型。

（3） CHAID 算法

与上述两个算法不同，此算法构建多分支的决策树。其分类的依据为卡方分析检验(Chi-square F test)，通过卡方检验来计算节点中的 P-value，来决定数据是否仍须进行分支。另外，CHAID 的目标变量和输入变量均适用于连续和分类类型。

（4） C5.0 算法

由 C4.5 演化而来，此算法的分类原理主要是利用信息衡量标准(Information Measure) 来构建决策树，并对每一个节点产生不同数目的分支来分割数据，直到数据无法分割为止。因此，C5.0 也是多分支的决策树算法。但其目标变量不适用于连续类型的变量，而输入变量则适用于连续或者分类类型的变量。

上述四种算法在分支依据、分支类型和适用的输入变量、目标变量的数据类型如表 8 所示：

表 8：决策树算法简介表

算法名称	分支依据	分支类型	输入变量类型	目标变量类型
C&RT	Gini 指数	二分支	连续或分类	连续或分类
QUEST	统计方法	二分支	连续或分类	分类
CHAID	卡方检验	多分支	连续或分类	连续或分类
C5.0	信息衡量标准	多分支	连续	分类

资料来源：申万宏源研究

此外，如第一章中所述，为了增强模型的稳定性，在接下来的不同决策树算法中均令 Bagging 次数为 10 来构建整体模型。

5.2 模型结果

在本小节中，分别采用上一小节中介绍的 C&RT、QUEST、CHAID、C5.0 方法对上市公司年报造假问题进行研究。首先，采用没有加入误分类损失函数的 C&RT 决策树对上市公司年报是否造假的判别问题进行研究。然后，再采用加入误分类损失函数的 C&RT 决策树对上述问题展开研究，并比较这两种情况下的 C&RT 决策树分类效果。判断样本数据的分类不平衡问题是否会对决策树的分类效果有所影响，以及加入误分类损失函数能否解决分类不平衡问题。

5.2.1 C&RT 结果

首先，不考虑样本数据的分类不平衡问题，直接构建决策树对上市公司年报造假问题进行研究。决策树的停止规则为父分支中的最小记录数为 2%，子分支中的最小记录数为 1%。先验概率基于训练集数据中的分类概率，目标的杂质测量标准为 Gini 指数。C&RT 决策树的准确率如下表所示：

表 9：C&RT 决策树准确率

分区	训练集	百分比率	测试集	百分比率
正确	5,324	97.83%	2,345	97.83%
错误	118	2.17%	52	2.17%
总计	5,442		2,397	

资料来源：申万宏源研究

由上表可知，C&RT 决策树在训练集和测试集中分类所得的准确率均为 97.83%，与神经网络模型和 SVM 模型相差不大。

另外，C&RT 决策树所得的混淆矩阵如表 10 所示：

表 10：C&RT 决策树混淆矩阵

训练集	0	1	测试集	0	1
0	5,324	0	0	2,345	0
1	118	0	1	52	0

资料来源：申万宏源研究，注：横坐标“0”、“1”分别表示预测结果为“非造假样本”、“造假样本”，纵坐标的“0”、“1”分别表示实际为“非造假样本”、“造假样本”。

根据表 10 中的结果可知，由于样本数据的分类不平衡问题，导致 C&RT 决策树在分类的时候完全无法判断出造假样本。与神经网络和 SVM 模型不同的是，决策树种可以使用误分类损失函数。在此，我们假设把造假样本判断为非造假样本所导致之损失为把非造假样本判断为造假样本所带来之损失的 50 倍。

在加入了误分类损失函数之后的 C&RT 决策树的准确性如下表所示：

表 11：加入误分类损失之后的 C&RT 决策树准确率

分区	训练集	百分比率	测试集	百分比率
正确	5,180	94.34%	2,219	94.51%
错误	311	5.66%	129	5.49%
总计	5,491		2,348	

资料来源：申万宏源研究

C&RT 决策树在训练集中的准确率为 94.34%，在测试集中的准确率为 94.51%。因此，加入误分类损失在一定程度上降低了分类的准确率。另一方面，加入误分类损失之后的 C&RT 决策树混淆矩阵如下表所示：

表 12：加入误分类损失之后的 C&RT 决策树混淆矩阵

训练集	0	1	测试集	0	1
0	5,157	212	0	2,208	92
1	99	23	1	37	11

资料来源：申万宏源研究，注：横坐标“0”、“1”分别表示预测结果为“非造假样本”、“造假样本”，纵坐标的“0”、“1”分别表示实际为“非造假样本”、“造假样本”。

由上表计算得相关评价指标的值如下：

表 13：加入误分类损失之后的 C&RT 决策树评价指标结果

	训练集	测试集	全样本
准确率	94.34%	94.51%	94.39%
精确度	9.79%	10.68%	10.06%
召回率	18.85%	22.92%	20.00%
F 值	12.89%	14.57%	13.39%

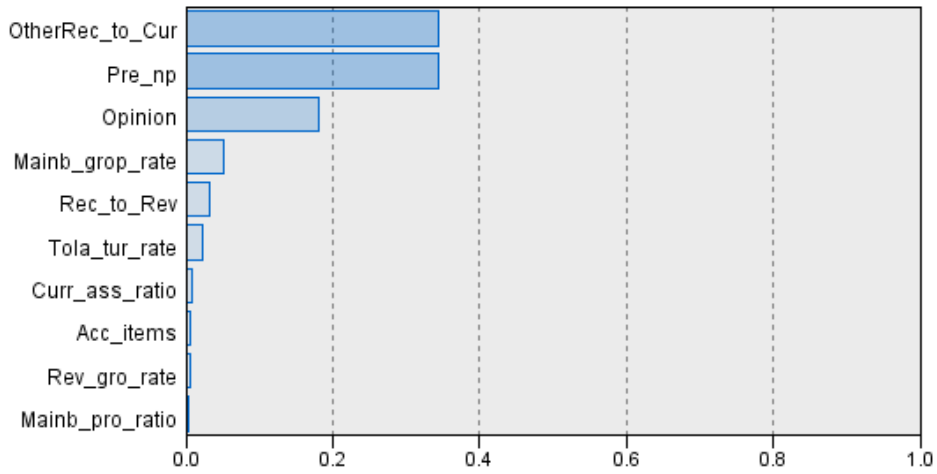
资料来源：申万宏源研究

由于加入了误分类损失函数，致使模型更看重是否能够识别出造假样本，提高了对造假样本分类准确率的重视程度。此时的 C&RT 决策树在全样本下的准确率为 94.39%，精确度为 10.06%，召回率 20.00%，F 值为 13.39%。因此，加入了误分类损失函数的 C&RT 决策树在一定程度上能够解决样本数据的分类不平衡问题，提高了对造假样本的判别能力。

由于在实际投资中，识别出财务造假的上市公司年报重要性更高。因此，在后续的 QUEST、CHAID、C5.0 这三个决策树算法中，均直接采用加入分类损失函数的决策树模型进行上市公司年报是否存在财务造假的判别研究。

此外，加入误分类损失函数之后的 C&RT 决策树中变量重要性情况如图 13 所示：

图 12 加入误分类损失函数之后的 C&RT 决策树变量重要性



资料来源：申万宏源研究

由上图可知，加入误分类损失函数之后的 C&RT 决策树中重要性排名在前五的变量分别为：OtherRec_to_Cur（其他应收款占流动资产比例）、Pre_np（造假前 1 年是否亏损）、Opinion（审计信息）、Mainb_grop_rate（主营业务毛利率）、Rec_to_Rev（应收账款/营业收入）。上述五个变量中，Opinion、OtherRec_to_Cur、Pre_np 也出现在神经网络模型、SVM 模型重要性排名前四的变量中。而相对于同行业而言，过高的主营业务毛利率和应收账款/营业收入比率也是识别出上市公司年报财务造假的重要指标。

5.2.2 QUEST 结果

同样地，设置误分类损失函数比例为 50:1，此 QUEST 决策树对上市公司年报是否存在财务造假问题的判别准确率如表 14 所示：

表 14：加入误分类损失函数的 QUEST 决策树准确率

分区	训练集	百分比率	测试集	百分比率
正确	4,952	90.18%	2,112	89.95%
错误	539	9.82%	236	10.05%
总计	5,491		2,348	

资料来源：申万宏源研究

由上表可知，无论是在训练集还是测试集，QUEST 决策树预测的准确性均比 C&RT 决策树预测的准确性低。另一方面，QUEST 决策树的混淆矩阵如表 15 所示：

表 15：加入误分类损失函数的 QUEST 决策树混淆矩阵

训练集	0	1	测试集	0	1
0	4,914	455	0	2,098	202
1	84	38	1	34	14

资料来源：申万宏源研究，注：横坐标“0”、“1”分别表示预测结果为“非造假样本”、“造假样本”，纵坐标的“0”、“1”分别表示实际为“非造假样本”、“造假样本”。

根据上述混淆矩阵，计算得 QUEST 决策树的相关评价指标如下表所示：

表 16：加入误分类损失之后的 QUEST 决策树评价指标结果

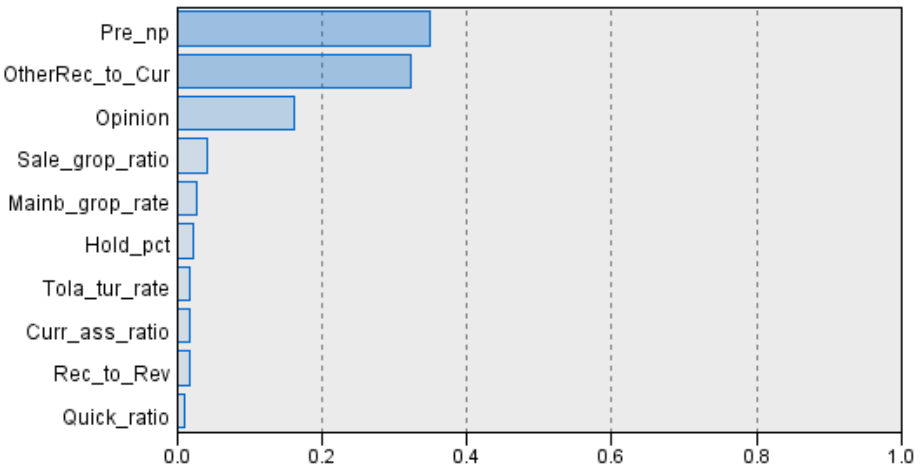
	训练集	测试集	全样本
准确率	90.18%	89.95%	90.11%
精确度	7.71%	6.48%	7.33%
召回率	31.15%	29.17%	30.59%
F 值	12.36%	10.61%	11.83%

资料来源：申万宏源研究

QUEST 决策树在全样本中的准确率为 90.11%，精确度为 7.33%，召回率 30.59%，F 值为 11.83%。因此，QUEST 决策树分类的准确率、精确度和 F 值均低于 C&RT 决策分类的准确率、精确度和 F 值。

QUEST 决策树中各个变量的重要性情况如下图所示：

图 13 加入误分类损失函数之后的 QUEST 决策树变量重要性



资料来源：申万宏源研究

由上图可知，加入误分类损失函数之后的 QUEST 决策树中重要性排名在前五的变量分别为 Pre_np（造假前 1 年是否亏损）、OtherRec_to_Cur（其他应收款占流动资产比例）、Opinion（审计信息）、Sale_grop_ratio（销售毛利率）、Mainb_grop_rate（主营业务毛利率）。上述五个变量与 C&RT 决策树中的前五大重要变量具有很高的重合度。由此，在一定程度上肯定了上述指标在上市公司年报财务造假预测中的重要性。

5.2.3 CHAID 结果

当父分支中的最小记录数小于等于 2%或者子分支中的最小记录数小于等于 1%时停止分支，加入 50 倍的误分类损失函数。此时加入误分类损失函数的 CHAID 决策树准确率如表 17 所示：

表 17：加入误分类损失函数的 CHAID 决策树准确率

分区	训练集	百分比率	测试集	百分比率
正确	5,138	93.57%	2,165	92.21%
错误	353	6.43%	183	7.79%
总计	5,491		2,348	

资料来源：申万宏源研究

由表 17 可知，加入误分类损失函数的 CHAID 决策树在训练集和测试集中的准确率分别为 93.57%、92.21%，稍微高于加入误分类损失函数的 QUEST 决策树准确率。而加入误分类损失函数的 CHAID 决策树混淆矩阵如下表所示：

表 18：加入误分类损失函数的 CHAID 决策树混淆矩阵

训练集	0	1	测试集	0	1
0	5,055	314	0	2,147	153
1	39	83	1	30	18

资料来源：申万宏源研究，注：横坐标“0”、“1”分别表示预测结果为“非造假样本”、“造假样本”，纵坐标的“0”、“1”分别表示实际为“非造假样本”、“造假样本”。

根据混淆矩阵计算所得的相关评价指标结果如下：

表 19：加入误分类损失之后的 CHAID 决策树评价指标结果

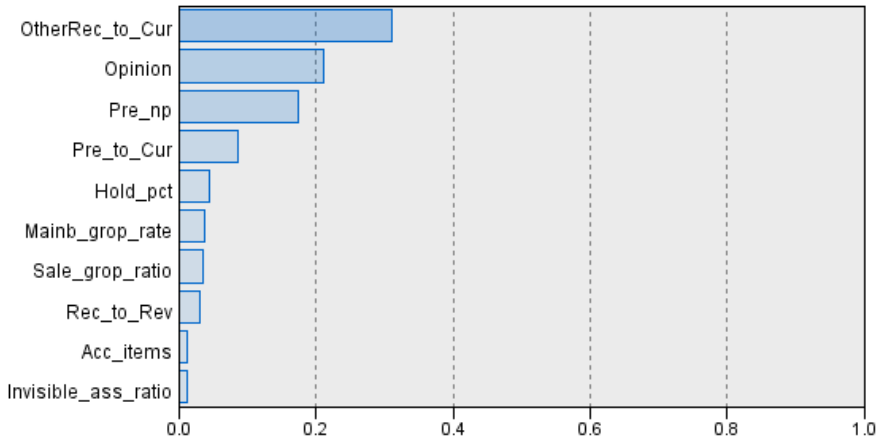
	训练集	测试集	全样本
准确率	93.57%	92.21%	93.16%
精确度	20.91%	10.53%	17.78%
召回率	68.03%	37.50%	59.41%
F 值	31.98%	16.44%	27.37%

资料来源：申万宏源研究

由上表可知，加入误分类损失函数的 CHAID 决策树在全样本中的准确率为 93.16%，精确度为 17.78%，召回率 59.41%，F 值为 27.37%。由此可知，加入误分类损失函数的 CHAID 决策树在识别造假样本的有效性上，远高于上述神经网络模型、SVM 模型和 C&RT 决策树、QUEST 决策树等在识别造假样本上的有效性。

CHAID 决策树中各个变量的重要性情况如下图所示：

图 14 加入误分类损失函数之后的 CHAID 决策树变量重要性



资料来源：申万宏源研究

由图 14 可知，加入误分类损失函数之后的 CHAID 决策树中重要性排名在前五的变量分别为 OtherRec_to_Cur（其他应收款占流动资产比例）、Opinion（审计信息）、Pre_np（造假前 1 年是否亏损）、Pre_to_Cur（预付款项占流动资产比例）、Hold_pct（前五大股东股权集中度）。其中除了 Pre_to_Cur 这个变量之外，其他四个变量均出现在前面几个模型的前几大重要性变量中，这进一步说明了上述变量在预测上市公司财务造假中的重要程度，也体现了模型的稳定性。而过高的预付款项占流动资产比例也是识别上市公司年报造假的一项重要指标，主要是因为当上市公司以虚构交易等方式来虚增利润时，需要以增加预付款项的方式来填补由于虚增利润导致的资金缺口，平衡资产负债表。

5.2.4 C5.0 结果

模型的相关设置与前面三个决策树相同，C5.0 决策树计算所得的准确率如下表所示：

表 20：加入误分类损失函数的 C5.0 决策树准确率

分区	训练集	百分比率	测试集	百分比率
正确	4,137	75.34%	1,685	71.76%
错误	1,354	24.66%	663	28.24%
总计	5,491		2,348	

资料来源：申万宏源研究

由上表可知，加入误分类损失函数的 C5.0 决策树在训练集和测试集中的准确率分别为 75.34%和 71.76%。该准确率均低于加入误分类损失函数的 C&RT 决策树、QUEST 决策树、CHAID 决策树的准确率。

另一方面，加入误分类损失函数的 C5.0 决策树混淆矩阵如下表所示：

表 21：加入误分类损失函数的 C5.0 决策树混淆矩阵

训练集	0	1	测试集	0	1
0	4,020	1,349	0	1,650	650
1	5	117	1	13	35

资料来源：申万宏源研究，注：横坐标“0”、“1”分别表示预测结果为“非造假样本”、“造假样本”，纵坐标的“0”、“1”分别表示实际为“非造假样本”、“造假样本”。

根据混淆矩阵计算所得的相关评价指标结果如下：

表 22：加入误分类损失之后的 C5.0 决策树评价指标结果

	训练集	测试集	全样本
准确率	75.34%	71.76%	74.27%
精确度	7.98%	5.11%	7.07%
召回率	95.90%	72.92%	89.41%
F 值	14.74%	9.55%	13.10%

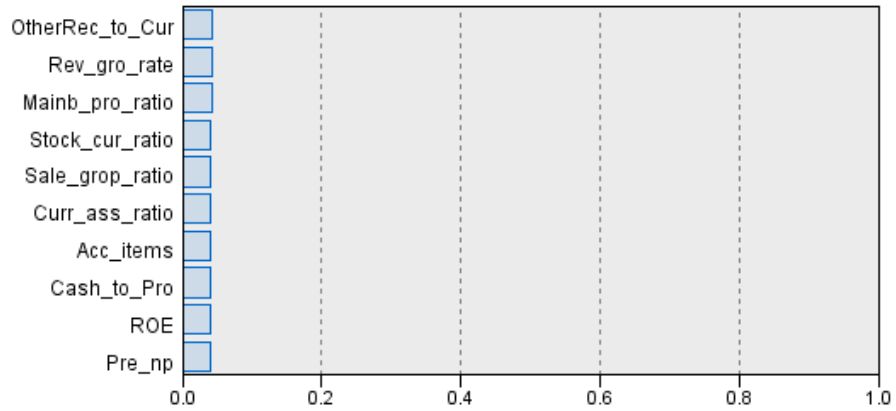
资料来源：申万宏源研究

由上表可知，加入误分类损失函数的 C5.0 决策树在全样本中的准确率为 74.27%，精确度为 7.07%，召回率 89.41%，F 值为 13.10%。由此可知，加入误分

类损失函数的 C5.0 决策树的召回率上远高于上述其他模型对造假样本的召回率。但是其判断的精确度为 7.07%，相对较低。

另外，加入误分类损失函数之后的 C5.0 决策树变量重要性情况如下图所示：

图 15 加入误分类损失函数之后的 C5.0 决策树变量重要性



资料来源：申万宏源研究

由图 15 可知，C5.0 决策树中重要性排名在前列的变量分别为：OtherRec_to_Cur（其他应收款占流动资产比例）、Rev_gro_rate（主营业务收入同比增长率）、Mainb_pro_ratio（主营业务利润占比）、Stock_cur_ratio（存货占流动资产的比率）、Sale_grop_ratio（销售毛利率）。

5.3 本章小结

本章前面几个小节的研究结果显示，加入误分类损失函数在四个决策树算法中均能提高决策树模型的召回率。四个算法在全样本中的准确率、精确度、召回率和 F 值如下表所示：

表 23：四个决策树算法的评价指标统计

决策树算法	准确率	精确度	召回率	F 值
C&RT	94.39%	10.06%	20.00%	13.39%
QUEST	90.11%	7.33%	30.59%	11.83%
CHAID	93.16%	17.78%	59.41%	27.37%
C5.0	74.27%	7.07%	89.41%	13.10%

资料来源：申万宏源研究

C&RT、C5.0、CHAID 在准确率上差别不大，均为 90%以上。C5.0 具有高达 89.41%的召回率，但是其精确度相对较低；而 CHAID 在精确度和 F 值中的表现为四个算法中最优。综合考虑模型预测的准确率、精确度、召回率和 F 值，加入误分类损失函数的 CHAID 决策树是相对较优的模型。其中，CHAID 决策树对上市公司年报是否造假判断的准确率为 93.16%，对造假样本的召回率为 59.41%，判断的精确度为 17.78%，F 值为 27.37%。

另外，综合考虑四个决策树模型中预测变量重要性结果，可以得出在分析上市公司年报是否造假的时候需要重点关注的 5 个指标分别为：

（1） Opinion，即审计师意见， 审计师接受委托对上市公司的经济活动、管理活动及其相关资料进行鉴证服务，提供鉴证报告。对外界而言，审计师的意见增强了财务信息的可信性，我们通过研究发现，当审计师给出“带强制事项段的无保留意见”或以上更为严重的审计意见时，其财务质量需要重点关注。

（2） Pre_np，即该公司在前一年中净利润是否亏损，由于中国股市存在 ST 制度，因此当某个公司前一年度财务亏损时，而当年的财务仍然亏损时，其有很大的动机对年报进行造假。因此，前一年净利润是否亏损也是判断一个上市公司是否会进行财务造假的重要指标。

（3） OtherRec_to_Cur，即其他应收款占流动资产比例，该比例相对比较高时，年报造假的可能性大。其原因主要为上市公司在进行年报造假的时候，大多数会通过各种方法来增加其他应收款，从而提高收入，虚增利润。

（4） Sale_grop_ratio，即销售毛利率，销售毛利率相对偏高的时候，年报造假的可能性也偏高。上市公司往往会通过关联交易等方法来减少相关成本，从而抬高销售毛利率。

（5） Pre_to_Cur，即预付款项占流动资产比例，预付款项占流动资产比率较大的年报造假可能性更高，这主要是因为上市公司会通过预付款项的形式来填补虚增收入的漏洞，以缓解虚增收入时带来的资金缺口压力，实现财务报表表面的正常平稳。

6、总结

本文根据财务造假常见的动机和手段，结合国内外学者的研究，构建了多个财务指标和非财务指标，分别采用数据挖掘中的神经网络、SVM 支持向量机、决策树（C&RT、QUEST、CHAID、C5.0），对上市公司年报造假判别问题进行研究。

由于上市公司造假案例相对于全体上市公司而言，数量占比较小，从而存在数据分类不平衡问题。该问题会导致在使用一些数据挖掘技术进行研究的时候，模型会整体向占大比例的分类数据拟合，从而忽视和无法判别小比例的分类数据。然而，在实际投资中，把造假公司判断为非造假公司所带来的成本远大于把非造假公司判断为造假公司。

在决策树中可以通过加入误分类损失函数的方式来让模型更加重视对造假样本的判别，因此，决策树比神经网络、SVM 支持向量机更适用于上市公司年报造假判别的研究，而决策树中的 CHAID 算法是四个算法中综合表现较优的算法。在全样本中，其对上市公司年报是否造假判别的准确率为 93.16%，召回率为 59.41%，判断的精确度为 17.78%，F 值为 27.37%。

最后，在判断上市公司年报是否造假的时候，需要重点关注的几项指标分别为：审计师意见、前一年是否亏损、其他应收款占流动资产比例、销售毛利率、预付款项占流动资产比例。审计师给出负面意见、前一年财务亏损、相对于同行业而言，比较高的其他应收款占流动资产比例、销售毛利率、预付款项占流动资产比例等，都是造假可能性高的上市公司年报所具有的特征。

信息披露

证券分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告,并对本报告的内容和观点负责。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

与公司有关的信息披露

本公司隶属于申万宏源证券有限公司。本公司经中国证券监督管理委员会核准，取得证券投资咨询业务许可，资格证书编号为：ZX0065。本公司关联机构在法律许可情况下可能持有或交易本报告提到的投资标的，还可能为或争取为这些标的提供投资银行服务。本公司在知晓范围内依法合规地履行披露义务。客户可通过 compliance@swsresearch.com 索取有关披露资料或登录 www.swsresearch.com 信息披露栏目查询从业人员资质情况、静默期安排及其他有关的信息披露。

机构销售团队联系人

上海	陈陶	021-23297221	18930809221	chentao@swsresearch.com
北京	李丹	010-66500610	18930809610	lidan@swsresearch.com
深圳	胡洁云	021-23297247	13916685683	hujy@swsresearch.com
海外	张思然	021-23297213	13636343555	zhangsr@swsresearch.com
综合	朱芳	021-23297233	18930809233	zhufang@swsresearch.com

法律声明