财务困境预测项目的Clementine数据流构建研究

李 盈 邓尚民 庄新磊

(山东理工大学科技信息研究所 山东淄博 255091)

【摘要】 文章以数据挖掘软件 Clementine 为平台,按照 CRISP-DM 的六个阶段对财务困境预测项目进行了数据流构建,利用 C5.0 算法生成的决策树建立预测模型,并对模型结果进行了分析。

【关键词】 财务困境预测 Clementine 数据流

Clementine 作为一个广泛应用于电子商务领域的软件,已经在客户流失分析、贷款欺诈等很多具体项目中得到了成功应用,但是在财务困境预测方面却少有涉及。数据挖掘软件具有很好的预测功能,如果能够在财务领域发挥其作用,将大大提高财务预测信息的质量。学者们对数据挖掘技术在财务困境预测中的研究仅仅局限在对具体技术优劣的比较,而没有着眼于整个项目进行研究。本研究将以 Clementine 为平台,研究财务困境预测项目的数据流构建,以期加速数据挖掘技术在财务领域的应用。

一、研究现状

陈晓(2000)是国内研究财务困境预测最早的学者之一,他采用逻辑回归方法进行建模,并基于国内资本市场的特殊性,在财务困境的定义和数据源选取方面做了许多工作。其将因财务状况异常而被特别处理(ST)的公司定义为陷入财务困境的公司。

吴俊杰(2006)介绍了数据挖掘技术的另一种方法——决策树方法,并比较了逻辑回归、神经网络和决策树在我国上市公司财务困境预测问题上的优劣,认为决策树在预测准确率、波动性及可解释性上具有综合优势。在数据源选取上,吴俊杰根据我国目前的实际情况做了改进,将企业被实施"退市风险警示"(*ST)而不是"特别处理"(ST)作为企业陷入财务困境的界定标准,这是因为被*ST 的企业从概念上与因为财务状况异常而被特别处理的企业基本一致,都排除了因非财务因素而导致被特别处理的情况,有利于提高模型预测的准确性。但他在利用决策树方法建模方面并没有展开介绍。

从以上研究可看出,学者们对财务困境预测的研究多集中在数据挖掘方法的选择和具体技术优劣的比较上。本文认为,财务困境预测项目的实现需要系统化的部署,而不只是对方法本身的研究。因此,本文以 Clementine 为平台,按照跨行业数据挖掘过程标准(CRISP-DM)来对上市公司的财务数据进行数据流构建。CRISP-DM 是当今数据挖掘领域最有影响的通用标准,其不只是对数据的组织或呈现,也不仅是数据分析和统计建模,而是一个从理解业务需求、寻求解决方案到接受实践检验的完整过程。这里引入 CRISP-DM 的目的就是将财务困境预测的数据挖掘过程视为一个系统来进行研究。

二、财务困境预测的数据流构建

Clementine 是能够有效地为企业改进决策的一个数据采集工作平台,提供了包括神经网络、决策树、聚类分析、关联分析、因子分析、回归分析等在内的丰富的数据挖掘模型,它通过节点的连接来完成整个数据挖掘过程,完全支持世界通行的 CRISP-DM,提供了从商业理解、数据理解、数据准备、建立模型、模型评估到结果部署的整个数据挖掘过程的项目管理功能。本文从以下六个阶段来对财务困境预测项目的数据流构建过程进行介绍。

1. 商业理解。这个阶段是决定数据挖掘项目的目标,并对项目实施的情况进行评估的阶段。

财务困境预测项目的目标就是建立一个预测模型,即利用已知的正常公司与陷入财务困境公司的财务数据,建立预测模型,分析两类公司在哪些财务指标上有显著差别,利用这个模型可以预测目标公司的财务数据满足什么条件时可视为有陷入财务困境的风险。

2. 数据理解。此阶段包括收集原始数据、描述数据及证实数据的质量,在数据理解的过程中需要清楚数据源是什么及数据源的特征。

项目的数据源是上市公司公开的财务数据,在财务困境标准界定这个问题上,本文沿用吴俊杰的做法,将企业被*ST 作为陷入财务困境的标准。然后找出 2007 年被*ST 的公司共 50 家,同时期的正常公司 1 298 家,找出这两类公司 2005 年的财务数据,存入 Excel 表格中。本文不选用 2006 年的上市公司财务数据建模,是因为上市公司 2006 年的财务数据公布时间在 2007 年 3 月份左右,与发布退市警示公告的时间过于接近,在这个基础上建模其准确率会被高估且实用性要差一些。

3. 数据准备。这个阶段是对用于挖掘的数据进行准备,包括选择、清理、重构、整合及格式化数据。

将存在 Excel 表格中的数据利用 Excel 导入节点读入数据流作为数据源,添加类型节点,在类型节点选项中设置"类

型”字段为输出字段,其他为输入字段。

数据源中*ST公司和正常公司的样本数量悬殊,是因为在现实中被退市警示的公司毕竟是少数,但是利用不平衡的数据建模会出现预测误差较大的情况,所以选择使用平衡节点来修正数据集的不平衡情况,使两类公司的样本数量达到均衡。在平衡节点模型选项中,将记录平衡指令的因子定为0.04,条件为“类型=正常”,即从Excel数据源样本中随机选取正常公司的4%,目的是使正常公司的数量与被退市警示公司的数量接近,这样Excel数据源中的正常公司数量与*ST公司的数量就达到了均衡。

4. 建模。数据准备就绪之后,最重要的就是选择合适的数据挖掘工具了。在Clementine中可以建立六种数据挖掘模型——分类模型、回归模型、时间序列模型、聚类模型、关联规则模型以及顺序规则模型。其中分类模型和回归模型主要是用来预测的,而回归模型用于对连续值的预测,不适合财务困境预测这种非连续值预测的情况。分类模型是用一些已经分类的数据来研究它们的特征,然后再根据这些特征对其他未经分类或新的数据进行预测,所以更适合该项目。随着数据挖掘技术的普及,分类方法中的神经网络和决策树逐渐应用到财务困境预测领域。

神经网络在建立预测模型时存在两个问题:一是其预测所依据的因素不明确,这样不利于对结果的分析;二是神经网络对测试数据可以进行相当正确的预测,但是对真实数据预测的准确性较差。不同于神经网络,决策树利用一系列的规则来得到一个类别或数值,所做的预测相对正确,且比神经网络容易理解。所以我们选择决策树方法建立数据挖掘模型。

选择Clementine中的C5.0算法生成的决策树建模,在模型选项中,选择“使用推进”和“交互验证”两个选项:①“使用推进”就是用C5.0算法中被称作“自举”的方法来提高模型的精确率,这种方法按序列建立多重模型,第一个模型以通常的方式建立,随后建立第二个模型,聚焦于另一个模型错误分类的记录,然后第三个模型聚焦于第二个模型的错误,最后应用整个模型集对数据样本进行分类,使用加权投票过程把分散的预测合并成综合预测。②选择决策树建模时,使用一组基于训练数据子集建立的模型,来估计基于全部数据建立的模型的准确性,选择“交互验证”可以解决因为数据集过小,不能将数据样本拆分成传统意义上的训练集和测试集的情况,在计算了准确性估计值后,用于交互验证的模型将被不再使用。

模型的结果如图1所示。按照输出字段“类型”将上市公司分为*ST和正常两类,根据能够带来最大信息收益的字段——资产负债率,将数据样本拆分为两个节点,节点0为根节点,节点1、节点2为叶子节点,从根节点到叶子节点为路径,一条路径就是一条规则。在图1中,资产负债率≤0.894和资产负债率>0.894就是“规则”,符合资产负债率≤0.894的上市公司全部为*ST公司,共有50家,否则为正常公司,共有55家。在建模过程中,那些对模型值没有显著贡献的字段则被剔除了。

决策树方法建立的模型中,资产负债率字段在区分陷入

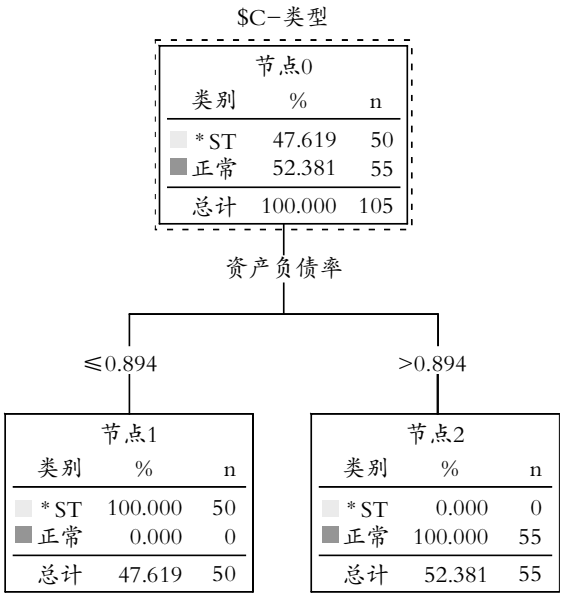

图1 决策树模型结果

财务困境公司和正常公司时的信息增益最大。如果目标公司资产负债率字段的值接近0.894或者因为某些原因使它的数值不具有说服力时,就需要参照其他字段来分类,因此本文添加了特征选择节点。该模型选项中的设置情况为:单个类别中记录的最大百分比设为90%;最大类别数作为记录的百分比设为95%;最小变异系数设为0.1。执行结果如图2所示。结果是按照对输出字段“类型”影响的程度来排列的,其中有重要影响的字段为资产负债率、每股净资产、净利率、现金流动比率和流动比率,速动比率为一般重要影响,其他字段对区分两类公司的影响不重要。

	秩	字段	类型	重要性	值
☑	1	#资产负债率	连续	★重要	1.0
☑	2	#每股净资产	连续	★重要	1.0
☑	3	#净利率	连续	★重要	1.0
☑	4	#现金流动比率	连续	★重要	0.995
☑	5	#流动比率	连续	★重要	0.974
☐	6	#速动比率	连续	⊕一般重要	0.928
☐	7	#净资产收益率	连续	⊙不重要	0.524
☐	8	#应收账款周转率	连续	⊙不重要	0.271
☐	9	#存货周转率	连续	⊙不重要	0.209

图2 特征选择模型的结果

5. 模型评估。对已经建立的决策树模型使用分析节点进行评估,结果如图3所示。其准确率达到了99.06%。

正确	105	99.06%
错误	1	0.94%
总计	106	100%

图3 决策树模型的评估结果

为了形象地表示各财务指标对两类公司的分类贡献程度，选用散点图节点，设置X轴为“类型”字段，Y轴分别为资产负债率（决策树节点选定的规则）、净利率（特征选择节点认为是重要字段）、速动比率（特征选择节点认为是一般重要字段）和应收账款周转率（特征选择节点认为是不重要字段），执行后的散点图分别见图4、图5、图6、图7。

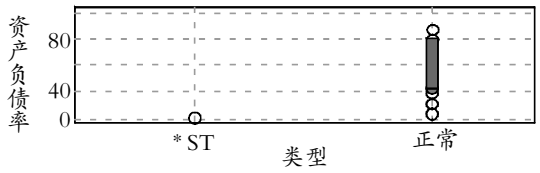

图4　资产负债率——类型散点图

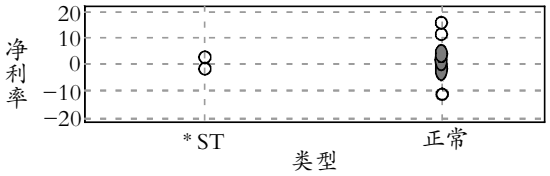

图5　净利率——类型散点图

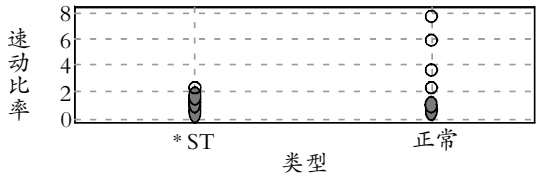

图6　速动比率——类型散点图

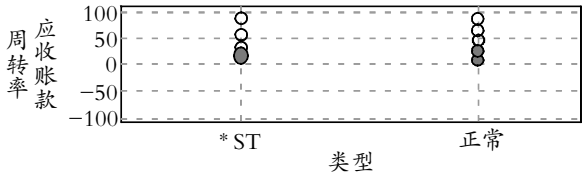

图7　应收账款周转率——类型散点图

通过对有代表性字段散点图的分析，可以形象地看出特征选择节点模型结果具有有效性。资产负债率和净利率属于两个重要字段，能够对类型字段进行较好区分，从图4和图5中可看出，正常公司90%以上都不同于*ST公司的数据；速动比率为一般重要性区分字段，从图6可以看出这个字段对两类公司的区分作用不是很明显；图7是不重要字段应收账款周转率的散点图，可以看到这个字段不能区分两类公司。另外，对比图4和图5，能够看出资产负债率字段比净利率字段的区分效果更好些，这表明决策树模型中选用资产负债率作为区分两类公司的标准是比较有说服力的。

6. 结果部署。至此，财务困境预测项目的数据流已经建立，如图8所示。结果部署阶段就是将建立的模型运用到实际中去，从而解决投资者的实际问题。数据挖掘工程师可以在已经建立的数据流中导入新的数据样本，结合遇到的新问题对数据流加以完善，对各个节点的设置进行修改调整，建立新的预测模型。

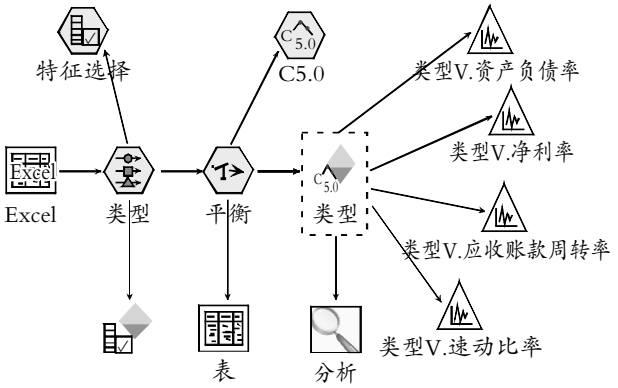

图8 财务困境预测项目的数据流程图

三、总结

本文建立了财务困境预测项目的数据流，其中平衡节点用来修正数据样本中被退市警示公司与正常公司数量不平衡的情况，以减少预测误差的出现。本文利用C5.0算法生成的决策树建立了财务困境预测模型，模型结果认为资产负债率字段在区分两类公司时的信息增益最大。本文利用分析节点对决策树模型的结果进行评估，评估表明决策树模型的正确率为99.06%。为了验证除资产负债率字段之外的其他字段的重要性程度，本文选择了特征选择节点，对所有字段在区分两类公司类型时的重要性进行排序。最后对几个有代表性的字段使用了散点图节点，以图的形式验证了决策树模型和特征选择模型的结果。

以往研究局限在具体数据挖掘方法在财务困境预测中的应用，本文引入跨行业数据挖掘标准流程（CRISP-DM）将财务困境预测项目视为系统，通过利用C5.0算法生成决策树建立预测模型。希望能够通过本文的研究加速数据挖掘软件Clementine在财务预测领域的应用进程。

本研究主要是财务困境预测数据流的建立，所以忽略了上市公司的外部影响因素（如政策影响或不可抗力因素）造成的被退市警示的情况，加之研究受到样本数据数量和质量的限制，模型结果的精确度会受到一定的影响。

主要参考文献

1. 陈晓，陈治鸿.中国上市公司的财务困境预测.中国会计与财务研究，2003；3

2. 吴俊杰.财务困境预测：数据挖掘方法的比较与运用.清华大学学报，2006；1

3. 吴世珍，柯大钢.我国上市公司财务危机预警研究.财会月刊（理论），2007；3

4. 宋素荣，于丽萍.上市公司财务危机预警的Logistic模型.财会月刊（理论），2006；10

5. 贺琼，郝汇.上市公司财务危机预警模型中变量体系的设计.财会月刊（理论），2007；2