基于粗糙集—决策树的上市公司财务预警

刘 澄(博士生导师) 胡巧红 孙 莹

(北京科技大学东凌经济管理学院 北京 100083)

【摘要】传统的财务预警研究往往把企业财务状况分成ST和非ST两类,过于笼统。为此本文首先运用聚类的方法把138家制造业上市公司分为财务状况健康、良好、一般、预警和危机5个层次,这使得对企业财务预警的研究更贴合实际,并且使实证研究结果更加准确。然后运用粗糙集中的变精度加权平均粗糙度来构造决策树的改进算法,对这些公司进行分类,进而提出公司财务状况预警的规则,这样生成的决策树财务预警规则防噪声能力更强,分类效果更好。

【关键词】决策树 粗糙集 财务预警

一、有关财务危机预测的研究方法

财务危机预测模型是由Beaver最早提出来的,之后许多预测方法被用于公司财务危机预测研究。20世纪60年代主要是Beaver和Altman分别采用单变量判别分析和多变量判别分析进行财务危机预警研究。20世纪80年代,Ohlson首先将Logistic模型应用于财务预警领域,20世纪90年代神经网络又被引入财务危机预测。20世纪80年代,Frydman等将决策树引入了财务预警研究中,决策树(DT)在解决分类问题上具有简单和易于理解的优点。

决策树是一种对大量数据集进行分类的非常有效的方法,通过决策树的构造模型,从大量信息中挖掘有效的数据,提取有价值的分类规则,从而获得有用的知识,帮助决策者准确预测。它的基本算法是贪心算法,采用自顶向下的递归方式构造决策树。

根据决策树增长的方法不同,学者们提出了很多经典的决策树算法。1986年J.R. Quinlan提出了决策树ID3算法,有人在此基础上提出了一些改进的SLIQ、SPRINT、CHAID等一些算法。这些算法运用也被运用到财务预警方面。姚靠华、陈晓红(2007)运用这些算法对我国上市公司的财务预警问题进行了研究。

1982年Z.Pawlak教授提出了粗糙集理论,运用粗糙集的方法可以对属性进行约简,把粗糙集的知识运用到决策树上,国内外学者提出了很多不同的建树方法并应用到很多领域。2001年赵卫东、李旗号运用粗糙集知识对决策树进行了优化,通过引入粗糙集理论中可分辨的概念给出一种方法,这种方法通过优化降低了树的高度。

2009年Iftikhar U. Sikder和Toshinori Munakata的基于粗糙集和决策树对低地震活动前兆因素的描述,他们运用粗糙集和决策树的方法,使用了信息增益和熵产生一系列规则,对地震进行预警。

运用决策树方法形成一系列规则,对训练数据集进行分类,然后根据形成的规则对训练数据集之外的数据进行分类,应用在财务领域,可以对财务进行预警。本文运用建造决策树的一种新方法,通过实证研究,对国内制造业上市公司进行财务预警分析。

二、基于变精度加权平均粗糙度的决策树生成算法

(一)对象聚类

系统聚类也称为层次聚类,是聚类分析中广泛应用的一种方法。聚类分析是建立在某种优化意义下,对样品或指标(变量)之间存在的相似性进行比较,将“相近似”的对象归并成类的一种方法。

本文使用SPSS16.0对138家制造业公司进行分类,聚类步骤如下:

1. 数据标准化。系统聚类首先要对各个原始数据进行一些相互比较运算,而各个原始数据往往由于量纲不同而影响这种比较和运算。因此,需要对原始数据进行必要的变换处理,以消除量纲不同造成的影响。

数据处理主要是对各个数据进行标准化,数据的标准化是将数据按比例缩放,使之落入一个小的特定区间,方法如下:

对于一个正向指标X_i,假定当它取值大于或者等于α时为最佳,此时,把它所有取值等于或者大于α的值标准化后取值为1;同理,假定当X_i的取值小于或者等于β时为最差,标准化后取值为0;取值为区间(β,α)的数据δ,标准化之后为:$(\delta-\beta)/(\alpha-\beta)$。

2. 计算聚类统计量。根据变换以后的数据计算得到聚类统计量。它用来表明各样品或变量间的关系相似或者近似程度。常用的统计量有距离和相似系数两大类。本文使用欧式距离计算聚类统计量。欧式距离计算方法如下:

假设每个样品有p个指标,用y_{ij}表示第i个样品的第j个指标,d_{ij}表示第i个样品与第j个样品之间的距离,欧式距离可表示为:

$$d_{ij}=\left[\sum_{k=1}^{p}(y_{ik}-y_{jk})^2\right]^{1/2} \quad (1)$$

3. 选择聚类方法。选择合适的聚类方法，将关系近似的样品或者变量聚为一类，关系不近似的加以区分。本文使用离差平方和法。假设将n个样品分成k类$G_1,\cdots,G_k$，用x_{it}表示类G_t中的第i个样品（x_{it}是p维向量），n_t表示G_t中样品个数，$\bar{x}_t$表示类G_t的重心，则在类G_t中的样品离差平方和是：

$$S_t=\sum_{i=1}^{n_t}(x_{it}-\bar{x}_t)'(x_{it}-\bar{x}_t) \quad (2)$$

整个类内平方和是：

$$S=\sum_{t=1}^{k}\sum_{i=1}^{n_t}(x_{it}-\bar{x}_t)'(x_{it}-\bar{x}_t)=\sum_{t=1}^{k}S_t \quad (3)$$

当k固定时，要选择使S达到最小的分类。Ward法就是找局部最优解的一个方法。其思想是先将n个样品各自成一类，然后每次缩小一类，每缩小一类离差平方和就要增大，选择使S增加最小的两类合并，直到所有的样品归为要求的类的个数为止。

本文运用聚类分析法，把所选择的国内制造业上市公司分为五类。根据各指标值的表现情况，第一类为财务状况健康的公司，第二类是财务状况良好的公司，第三类是财务状况一般的公司，第四类为财务状况预警的公司，第五类为财务状况危机的公司。

（二）确定最优分支属性

构造决策树过程，分支属性用来确定树的非叶结点，树的每一次生长都要确定一个分支属性，所以说分支属性的选择至关重要，直接影响分类的质量。

定义1：加权平均粗糙度：

$$\gamma_{R_i}=1-\sum_{j=1}^{m}\omega_j\mu_{R_i} \quad (4)$$

其中：β为分类误差，它的取值范围是[0,0.5]，$\mu_{R_i}(X_j)=|\underline{R}_iX_j|/|\overline{R}_iX_j$，$\omega_j=|X_j|/|U|$，$R_i$表示第i个条件属性，m是决策属性等价类的个数，j表示决策属性的第j个等价类，U表示论域，X_j表示决策属性的第j个等价类集合。

定义2：变精度加权平均粗糙度：

$$\gamma_{R_i}^{\beta}=1-\sum_{j=1}^{m}\omega_j\mu_{R_i}^{\beta}x_j \quad (5)$$

其中：$\mu_{R_i}^{\beta}(X_j)=|\underline{R}_i^{\beta}X_j|/|\overline{R}_i^{\beta}X_j|$，$\omega_j=|X_j|/|U|$，$R_i$表示第i个条件属性，j表示决策属性的第j个等价类，X_j表示决策属性第j个等价类的集合，m是决策属性等价类的个数。$\underline{R}_i^{\beta}X_j$称为$X_j$的β下近似，$\overline{R}_i^{\beta}X_j$称为$X_j$的β上近似。$R_{ip}$表示第i个条件属性的第p个等价类，$r_{ip,n}\in R_{ip}$，则有：

$$\underline{R}_i^{\beta}X_j=\bigcup_{p=1}\{r_{ip,n}\in R_{ip}\,|\,|X_j\cap R_{ip}|/|R_{ip}|\geq 1-\beta\} \quad (6)$$

$$\overline{R}_i^{\beta}X_j=\bigcup_{p=1}\{r_{ip,n}\in R_{ip}\,|\,|X_j\cap R_{ip}|/|R_{ip}|\geq \beta\} \quad (7)$$

现实数据库中不可避免地存在很多噪声数据，使用变精度近似精度可以克服噪声数据对精确性的影响，在一定程度上消除噪声数据对刻画精度的影响。

$\gamma_{R_i}^{\beta}$的取值范围是[0,1]，它越小则反映第i个属性包含的近似确定性越大。于是，在决策树生长过程中，每次选择值最小的属性作为分支结点。

（三）基于变精度加权平均粗糙度构造决策树算法

决策树自顶向下生长，每次生长都选择变精度加权平均粗糙度值最小的属性作为树的分支属性。输入决策表和分类误差β，即可输出一棵决策树。算法步骤如下：

步骤1：根据输入的决策表计算每一个条件属性的变精度加权平均粗糙度，并比较它们的大小。

步骤2：选择变精度加权平均粗糙度最小的属性ψ作为决策树分支的属性。

步骤3：用选择的属性ψ去划分训练集，相应的该属性的每一个取值产生一个分支（子表），这样训练集被划分为若干小的决策表。

步骤4：若子表中属于某一类别实例个数占表中总实例个数大于等于（1-β）或表中没有可选的属性，则以该子表中占多数的实例类别标识该节点，并作为叶子结点；否则，将子表中的条件属性去掉已选划分属性ψ，重复以上步骤。

步骤5：返回。

算法步骤比较简单，决策树使用递归算法，对训练集划分，可以得到一个局部最优解。

三、仿真运算

1. 样本描述。本文随机选取2010年国内1 192家制造业上市公司中的138家公司的财务数据作为样本数据，并对数据进行标准化。然后，把每个条件属性按照标准化后的数值划分为5个等价类，分别为：[0,0.2）表示财务状况差，[0.2,0.4）表示财务状况较差，[0.4,0.6）表示财务状况中等，[0.6,0.8）表示财务状况较好，[0.8,1]表示财务状况好。由此，每个条件属性都有5个等价类。

本文选择了影响公司财务状况的五个方面的9个指标对企业进行评价，这五个方面分别是：盈利能力指标、现金流量指标、偿债能力指标、成长能力指标以及营运能力指标。具体的指标选择见下表：

选择的财务指标

盈利能力	R_1：净资产收益率	净利润/净资产
	R_2：总资产报酬率	EBIT/总资产
现金流量	R_3：现金流动负债率	年经营现金净流量/年末流动负债
偿债能力	R_4：流动比率	流动资产/流动负债
	R_5：资产负债率	总负债/总资产
成长能力	R_6：营业收入增长率	（当年营业收入-上年营业收入）/上年营业收入
	R_7：营业利润增长率	（当年营业利润-上年营业利润）/上年营业利润
营运能力	R_8：存货周转率增长率	（当年存货周转率-上年存货周转率）/上年存货周转率
	R_9：应收账款周转率增长率	（当年应收账款周转率-上年应收账款周转率）/上年应收账款周转率

2. 仿真运算。仿真运算软件：MatlabR2009a，SPSS16.0。分类误差β=0.25。为了方便起见，我们把138家公司按1～138进行了编号。

首先，运用系统聚类的方法把138家制造业上市公司分为5个等价类：分别以X_1、X_2、X_3、X_4、X_5，分别代表财务状况危

机、预警、一般、良好和健康的公司集合。

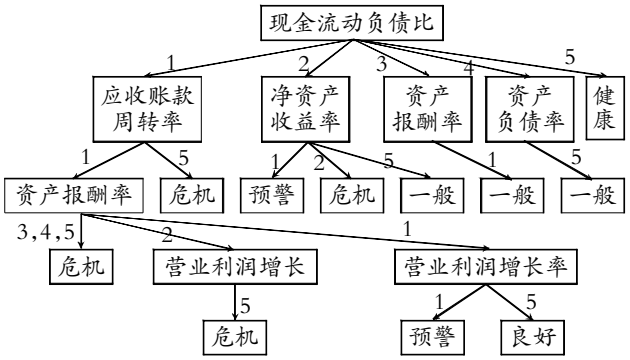
以标准化后的数据表计算每个条件属性的变精度加权平均粗糙度，分别为 $\gamma_{R_1}^{0.25}=1$，$\gamma_{R_2}^{0.25}=1$，$\gamma_{R_3}^{0.25}=0.879\ 2$，$\gamma_{R_4}^{0.25}=1$，$\gamma_{R_5}^{0.25}=1$，$\gamma_{R_6}^{0.25}=1$，$\gamma_{R_7}^{0.25}=1$，$\gamma_{R_8}^{0.25}=1$，$\gamma_{R_9}^{0.25}=0.976\ 9$。此时，条件属性$R_3$的变精度加权平均粗糙度为最小，选择$R_3$，也即现金流动负债率为该树的根结点，根据现金流动负债率的五个等价类生成5个子表。

若子表中属于某一类别实例个数占表中总实例个数比例大于等于0.75或者表中没有可选的属性，则以该子表中占多数的实例类别标识该节点，并作为叶子结点；否则，将子表中的条件属性去掉已选划分的条件属性R_3，重复以上步骤。

四、运算结果

以现金流动负债比作为树的根结点，把训练集分成了五类，其中现金流动负债比在区间[0.8,1]之间的公司有22家，而这22家中有19家是属于财务状况表现健康的公司，19/22=0.86，大于1-β。因此，现金流动负债比在区间[0.8,1]之间的等价类财务状况为健康。据此，我们可以认为训练集以外的公司现金流动负债比在区间[0.8,1]的为健康。而且，得出的决策树把大部分被ST的公司分到财务危机的一类，总体分类准确率比较高。

树中1表示区间[0,0.2)，2表示区间[0.2,0.4)，3表示区间[0.4,0.6)，4表示区间[0.6,0.8)，5表示区间[0.8,1]。决策树生成结果如下图：



根据决策树形成规则，共计11条规则：

规则1：现金流动负债比“差”，应收账款增长率“差”，资产报酬率“差”，营业利润增长率“差”→财务状况“预警”。

规则2：现金流动负债比“差”，应收账款增长率“差”，资产报酬率“差”，营业利润增长率“好”→财务状况“良好”。

规则3：现金流动负债比“差”，应收账款增长率“差”，资产报酬率“较差”，营业利润增长率“好”→财务状况“危机”。

规则4：现金流动负债比“差”，应收账款增长率“差”，资产报酬率“中等”或“较好”或“好”→财务状况“危机”。

规则5：现金流动负债比“差”且应收账款增长率“好”→财务状况“危机”。

规则6：现金流动负债比“较差”且净资产收益率“差”→财务状况“预警”

规则7：现金流动负债比“较差”且净资产收益率“较差”→财务状况“危机”。

规则8：现金流动负债比“较差”且净资产收益率“好”→财务状况“一般”。

规则9：现金流动负债比“中等”且资产报酬率“差”→财务状况“一般”。

规则10：现金流动负债比“较好”且资产负债率“好”→财务状况“一般”。

规则11：现金流动负债比“好”→财务状况“健康”。

五、结论

由于存在数据的可得性问题，传统的财务预警研究一般都是把企业财务状况分成ST和非ST两类，分别代表有预警和安全两种。本文运用聚类方法把企业财务状况分成健康、良好、一般、预警和危机5个层次，使得研究过程与企业实际状况更加契合。在此基础上，再以变精度加权平均粗糙度作为构造决策树的分支属性选择标准进行财务预警研究。

通过实证研究可以看出，决策树形成11条规则，并且构造出的决策树把大多数ST公司分到财务状况表现危机的一类，分类效果比较好。根据决策树形成的一系列规则，可以对训练集以外公司的财务状况进行预警。由于引入了变精度加权平均粗糙度，生成的决策树有效弱化了少数实例对决策树造成的不良影响，虽然决策中存在一定的误差，但决策树总体分类是比较好的，最终生成的决策树也比较符合实际，分类精度也比较高，且有效地处理了噪声数据。

主要参考文献

1. Beaver W H. Financial ratios as predictors of failure, empirical research in accounting: Selected studies . Journal of Accounting Research, 1966

2. Altman E I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy . The Journal of Finance, 1968;23

3. Ohlson J A. Financial ratios and probabilistic prediction of bankruptcy .Journal of Accounting Research, 1980;18

4. Tam K, Kiang M. Predicting bank failures: A neural network approach.Applied Artificial Intelligence, 1992;8

5. Frydman H, Altman E I, Kao D. Introducing recursive partitioning for financial classification: the case of financial distress . Journal of Finance, 1985;40

6. J. R. Quinlan Induction of Decision Tree .Machine Learning, 1986;1

7. 姚靠华，蒋艳辉. 基于决策树的财务预警. 系统工程，2005;23

8. 陈晓红，易松青.基于CHAID方法的中小企业上市公司财务预警研究. 经济与管理研究，2007;3

9. 赵卫东，李旗号.粗集在决策树优化中的应用.系统工程学报，2001;16

10. Iftikhar U. Sikder, Toshinori Munakata. Application of rough set and decision tree for characterization of premonitory factors of low seismic activity.Expert Systems with Applications, 2009;36