信用评分模型优化以及保险公司发现利基市场的实证研究

刘扬，刘伟江

（吉林大学商学院，吉林 长春 130012）

摘要： 运用数据挖掘分类技术于信用评分问题包括一个建立模型以及选择最优模型的过程。在这个过程中，需要有效地利用已知的训练样本，运用不同的分类算法建立结构适度的模型；以及评价各种算法建立的模型,选择其中最优的应用到实际问题中。本文对此优化过程进行了一个实证研究。一方面，提出了对解决实际问题具有参考价值的信用评分模型优化的两阶段模式，首先运用数据挖掘的增量分析法，确定不同算法的参数及其适量的训练样本，然后运用ROC曲线评价比较不同模型的准确率。另一方面，将这个优化模式应用于实际问题，运用优化的信用评分模型实现了某信用保险公司利基市场的发现。

关键词： 信用评分模型；模型评价；分类技术；利基市场

中图分类号: F830 **文献标识码: A**

1 引 言

信用评分是指利用历史样本数据建立模型，将客户信用风险水平用数值或类别值表示的定量方法。这种方法在国外消费信贷领域的应用起始于20世纪50年代，被银行及信用卡公司用来支持优化其信用决策过程，对于获得和保持竞争优势，发现和填补尚未受到充分服务的利基市场，起到至关重要的作用。此外，信用评分模型在企业信用评估以及公司财务危机预警等领域也具有重要应用价值。

近年来，国内外的研究开始将数据挖掘的分类预测技术运用于信用评分模型的建立。从数据挖掘的角度来看，信用评分模型的建立是一个从分析问题、收集数据、到建立模型、评价模型和实施应用的多阶段过程。其中，在建立模型和评价模型阶段如何选出最优最适用的模型是最关键也是最难以解决的问题。

在建立模型阶段，除了传统的线性判别分析和逻辑回归等多元统计分析算法之外，包括分类树、回归树、神经网络、遗传算法和K－最近邻法在内的诸多人工智能和非线性统计等领域的分类算法也已经成功地应用于信用评分模型的建立[1-9]。各种算法设定的参数决定模型的结构，并影响模型预测的准确性。但是在以往的研究中，如何确定模型参数没有得到充分关注，只是通过反复的试验寻找效果可行的模型。Weiss和Indurkhya[10]提出的数据挖掘增量分析法，运用学习曲线确定最优的模型参数和最适合的训练样本数，是一个值得尝试的方法。但至今这方面的研究仍很少，只有Galindo和Tamayo[6]的研究采用了增量分析法，科学地确定了适合当前问题和样本数量的模型结构。

在评价模型阶段，需要对运用不同建模算法所产生的多个模型进行评价，从中选择最优的模型。国内外各种模型的比较研究中运用了各种评价方法，首先最简单的是利用检验错误率衡量模型的准确性[4-8]，并比较模型的两种错误率（α错误和β错误）。当α错误和β错误的成本代价

不相等时，检验错误率过于简单片面。其次， Joos等[9]的研究利用成本方程计算出模型判断错误所导致的预期损失，比较不同的α和β错误成本比率下模型的预期损失，Nayak和Turvey[11]针对一个具体贷款问题研究了两种错误成本的估算，并运用计算得出的错误成本作为评价模型的标准。虽然，在理论上成本方程可以反应模型在应用中的效果，但是在实际问题上很难给出两种错误成本的准确估计，因此限制了其广泛应用。另外一个实用性强、效果好的评价方法是ROC曲线[12,13]。它可以给出不同分割值下两种错误率的连续变化关系，既避免实际计算的困难，又准确地反应了模型总体的预测能力。Hand等[14,15]的研究采用了ROC方法。目前国内的信用评分模型比较研究中还没有运用ROC方法。

本文对信用评分模型的建立及评价方法进行了系统的研究。首先提出了一个两阶段的信用评分模型优化模式。然后以一个保险公司的信用保险审批问题为例，运用所提出的优化模式给出了模型的建立及评价过程:（1）模型建立阶段,建立了两百多个模型，采用学习曲线进行增量分析，以优化模型的参数设定；（2）模型评价阶段，运用ROC曲线对五种常用的信用评分模型技术的预测能力作了比较分析。最后阐述如何应用优化的评分模型解决了信用保险公司发现利基市场的实际问题。

2 两阶段信用评分模型优化模式

2.1 模型建立阶段——增量分析与学习曲线

评价信用评分模型性能的基本指标是模型的错误率,通常利用检验样本得出，也称为检验错误率。本文利用学习曲线进行的增量分析是从训练样本量的变化和模型结构复杂度的变化两个方面，来观察模型检验错误率随之的变化，并以此确定模型适合的参数（参数决定模型复杂度）和判断训练样本是否充足。学习曲线将模型检验错误率的变化趋势用曲线表示出来。理论上，样本量和模型复杂度增量分析的学习曲线变化趋势如图1－a，1－b所示。

样本量（sample size）增量学习曲线（见图1－a）：评分模型的建立是一个利用训练样本的学习过程,学习的结果得出数据的分类模式。一般来说，随着训练样本量的增多，所获得的模型检验错误率会降低。然而，降低的速度和程度取决于具体的问题及其样本数据。有些数据所表示的模式比较复杂，当增加样本量时，错误率会连续地大幅度降低。有些数据所表示的模式比较简单，样本量增加到一定程度时，错误率不会继续降低，而是保持在一个水平，意味着样本数量已经足够描述其所代表的模式。因此，观察样本量增量学习曲线可以判断当前训练样本是否充足。

模型复杂度（model size或 model complexity）增量学习曲线（见图1－b）：评分模型的结构形式取决于所采用的算法。例如，决策树、人工神经网络算法建立的评分模型，其模型结构的复杂度为：决策树的节点越多、神经网络的隐含层神经元数目越多，模型结构的复杂度越大。模型结构可以由算法设置的参数调整。模型结构的复杂度不足时，模型不能很好地拟和数据，导致模型的检验错误率比较高，这说明模型的能力不足以描述当前数据所代表的复杂模式。可是如果模型结构过于复杂（过多的树节点，或者过多的隐含层神经元）就会引起“过学习”现象，反而导致模型检验错误率的增加。因此，通过复杂度增量学习曲线可以找到最适合当前数据和问题的模型结构。对于模型复杂度和模型准确率的关系Galindo和Tamayo给出了形象的描述[6]（见图2）。

对于某些算法，模型结构是否适度有时还受训练样本量的影响。模型的错误率随着样本数量的增加而不再降低，这时如果增加模型复杂度还可以继续降低错误率，说明数据增多后，其代表的模式更复杂，模型复杂度也需相应增加。

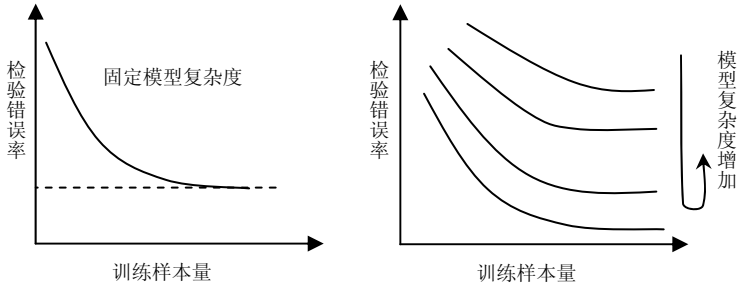

图 1-a.模型增量分析－样本量增量学习曲线

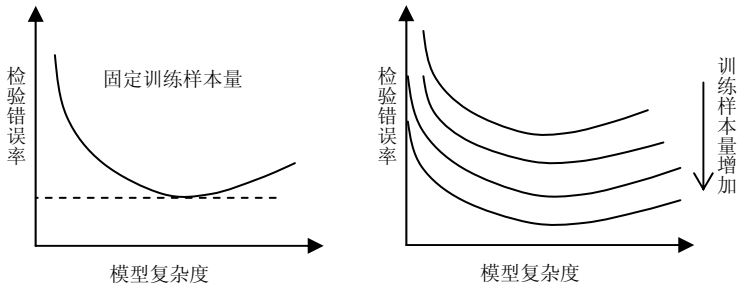

图 1-b.模型增量分析－复杂度增量学习曲线

说明：固定模型复杂度而增加样本量，错误率先降低，然后渐近某一水平（如图1-a）。固定样本量而增加模型的复杂度，开始会降低模型的错误率，到达最低点之后，错误率反而重新上升（如图1-b）。

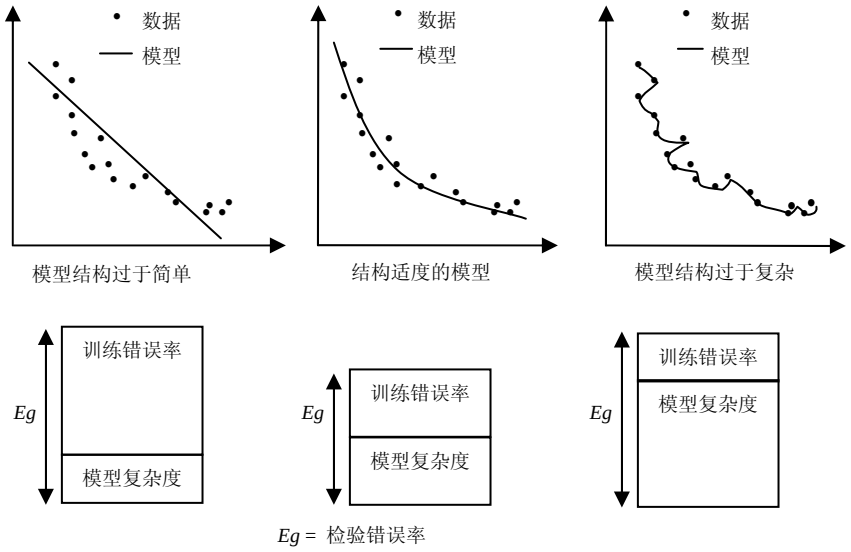

图 2　模型复杂度与模型准确率关系示例

在实际中，可能会有以下情况发生 ：（1）如果在某一种模型结构下样本量增量学习曲线上的错误率仍有下降趋势，则意味着样本量不够充足。在条件允许的情况下，应该补充采集更多的样本，从而有效地提高模型准确率。（2）如果在任何一种模型结构下的学习曲线都显示了错误率没有再降低的可能，则可以找到最优的模型结构及其适合的样本数量，从而用最少的样本和最合适的模型结构建立最优模型。这也就意味着以最低的成本建立最优的模型，因为减少样本数量既可以降低样本采集成本，又节省了模型训练时间。

2.2 模型评价阶段——比较分析与ROC曲线

模型的检验错误率实际上包含两种错误率，即α错误率和β错误率，分别代表信用差的样本被错误判断的比率和信用好的样本被错误判断的比率。运用模型检验错误率作为评价模型的标准意味着一个假定，即两种错误导致的损失相等。这个假定显然过于简单，不符合实际。因为α错误的损失一般大于β错误的损失。理论上，如果两种错误的损失可以计算，那么可以用下面公式计算出由于模型的错判而导致的损失的期望值。该期望值称为预期成本，可以作为评价信用评分模型的标准：

$$\text{预期成本} = P_b \cdot C_\alpha \cdot \alpha\text{错误率} + P_g \cdot C_\beta \cdot \beta\text{错误率},$$

其中 P_g, P_b = 总体中信用好坏的比率

C_α, C_β= α错误和β错误的损失成本

在实际中，评分模型的结果通常是连续的评分值，需要设定一个分割值，在分割值以上的判定为信用好的，在分割值以下的判定为信用差的。使用不同的分割值判别信用好坏，可以得出不同组合的α和β错误率，从而得出如图3所示的ROC曲线①。其纵坐标代表在分割值以上的信用好样本的比率，横坐标代表在分割值以上的信用差样本的比率。运用ROC曲线可以描述在不同分割值下α错误率和β错误率之间此消彼涨的连续变化的关系。

不失其比较作用，预期成本可变换为：

$$1 + (C_\alpha P_b)/(C_\beta P_g) \cdot \alpha\text{错误率} - (1 - \beta\text{错误率})$$

如果可以确定$(C_\alpha P_b)/(C_\beta P_g)$，就可以确定使预期成本最低的分割值，它对应于以$(C_\alpha P_b)/(C_\beta P_g)$为斜率的直线与ROC曲线的切点。在实际问题中，两种错误导致的损失的比率(C_α/C_β)以及总体中信用好坏的比率(P_b/P_g)很难准确确定，分割值经常是根据实际需要变化调整

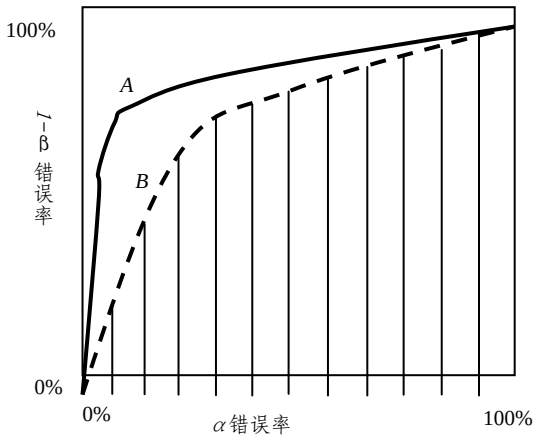

图 3.运用ROC曲线比较信用评分模型（模型*A*优于模型*B*）

的。因此，一种灵活地评价模型准确率的标准是ROC曲线下面积（area under curve, 简称AUC），AUC指ROC曲线向下和坐标轴之间的阴影部分所代表的面积。AUC越大，其所代表的模型总体准确率越高。AUC实现了从总体上衡量和比较模型在不同分割值下的两种错误率的发生情况。

本文采用上面介绍的方法，将信用评分模型的优化过程分为两个阶段：在建立模型阶段，运用学习曲线进行增量分析，以确定样本数量是否充足，决定模型结构的适当复杂程度；在评价模型阶段，运用ROC曲线下面积进行各种模型的比较，以选择最优模型。

3 信用评估实际问题

本文以德国某信用保险公司实际信用评估问题为例。保险公司与卖方企业之间针对应收帐款达成一种合约，如果卖方企业在售出货物后不能收回应收款项，保险公司将补偿其损失。为此，信用保险公司要分析买方企业的信用风险水平，也即它们在未来拖欠货款的可能性，从而决定是否批准卖方企业的保险申请。该保险公司从某信用调查中介机构获得的买方企业基本信息包含：企业成立年限、注册资本、企业类型、职工人数、所在行业、连续三年的销售额、企业信用综合指数、企业订货情况指数、企业发展状况指数、企业历史支付情况指数、固定资产是否抵押以及抵押比例，共14个变量。根据这些基本信息，保险公司的专家将企业信用状况分为表1所示的四个级别。

表 1.企业信用等级分类

A	企业信用状况很好，几乎不存在信用风险.
B	企业信用状况良好，信用风险比较低.
C	企业信用状况中等，存在一定的信用风险.
D	企业信用状况很差，存在很高的信用风险.

专家认为，针对*A/B* 类企业的保险申请如果小于一定的额度，通常应予以批准，针对*D* 类企业的申请应予以拒绝。而针对 *C* 类企业的申请应视具体情况而定。在总体样本数据中17%以上的企业属于 *C* 类，如果能有效的从 *C* 类企业中确定出信用良好的企业，就能找到以前被忽略的利基/缝隙市场(Niche Market)，从而增加信用保险收入。因此这个信用决策的关键点和难点就是如何判别 *C* 类企业的信用状况。

为建立两类问题的信用评分模型，将企业分为信用好（*A/B* 类）和信用差（*C/D*类）两类。经反复试验，这样的分类可以得出有效的评分模型以对 *C* 类企业的信用状况作出进一步细分（见模型实际应用部分的讨论）。从已知总体样本中预留出检验样本2，从其余样本中随机抽取出相互独立的训练样本1和检验样本1，用于评分模型的增量分析，然后再重新随机抽取训练样本2，用于最终评分模型的建立（见表2）。这样可以保证预留出的检验样本2的独立性，从而保证最终模型检验的有效性。

4 模型优化过程

本文的研究运用以下介绍的三种常用的分类预测算法，在模型建立阶段对其进行增量分析。在模型评价阶段，还加入两种在实际中经常采用的信用建模技术－逻辑回归 (Logistic Regression，简称LR)和线性判别分析（linear discriminant analysis, 简称LDA），以求扩大比

较的范围，体现比较结果的实际意义。

表 2. 建立模型的样本

	样本	样本量		样本用途
		“好”	“差”	
模型建立阶段的增量分析	训练样本 1	5000	5000	建立增量分析模型
	检验样本 1	2000	2000	检验增量分析模型 绘制学习曲线
模型评价阶段的比较分析	训练样本 2	X	X	建立最终模型
	检验样本 2	3820	3820	检验比较最终模型 绘制 ROC 曲线

说明：训练样本2的样本量 X 由增量分析结果决定，不同的算法有所不同。

误差反向传播神经网络 (Multi-layer perceptron with back-propagation，简称 MLP)：MLP 由输入层、隐含层和输出层的相互连结的神经元组成。通过样本数据对模型的多次反复训练，沿输出结果与实际结果的误差的负梯度方向修正各层神经元的权值和阈值，如此反复，直至网络全局误差平方和达到预期精度。本文采用了三层结构(输入层、一个隐含层和输出层)。激活函数为 sigmoid 函数，学习率初始设为 0.3，训练中若网络不收敛，则自动调低学习率重新训练。

k-最近邻法 (k-nearest-neighbors，简称 kNN)：存储一个由已知类别的训练样本组成的实例集合，当判别一个新实例的类别时，与其最相近的 k 个实例中多数属于哪个类别，新实例就被分到哪一个类别。为此，要给定一个计算方法用以衡量实例之间的相近程度。本文运用欧几里德距离函数定义两个实例的相近程度。

回归树（简称 M5）：M5 算法采用了一种和分类树类似的非返回跟踪的分割方法，将样本集递归分割成不相交的子集，分割准则旨在极大化分割子集包含样本的相似性。和一般分类树算法的不同之处在于，用该方法建立的树的叶节点不是离散的类别值，而是线性回归模型模拟的连续值。

4.1 第一阶段——增量分析

分别对MLP、kNN 和 M5 算法做了样本量增量分析和模型复杂度增量分析，结果见图4－图6。图中的样本数目100％＝10000。

MLP模型的复杂度由隐含层神经元数目 H 决定。从图4－a中可知，由于神经网络的不稳定性，模型错误率不总是随着训练样本数目的增多而单调下降。从总的趋势来看，错误率不会因为增加训练样本量而获得进一步降低。因此，在第二阶段建立MLP模型时，不需要增加训练样本的数量。从图4－b中可知，对于100％的样本量，合适的模型结构为 $H = 3$。多于3个神经元的模型错误率开始上升。

kNN模型复杂度由相邻的样本数 k 决定。从图5中可知，模型错误率一直随着训练样本数目的增多而单调下降。再增加一倍的训练样本量，错误率得到进一步降低。因此，在第二阶段建立 kNN 模型时使用了全部可用的训练样本（20360个）。选择使模型检验错误率最低的模

型参数 $k = 7$。

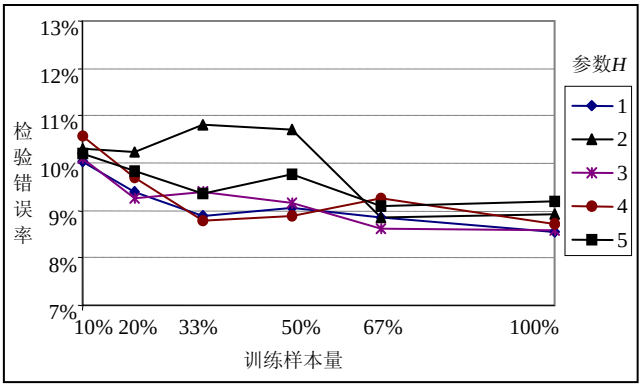

图 4-a: MLP模型的样本量增量分析

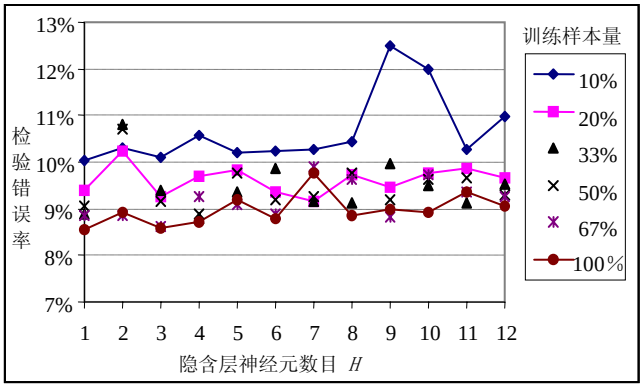

图 4-b: MLP模型的复杂度增量分析

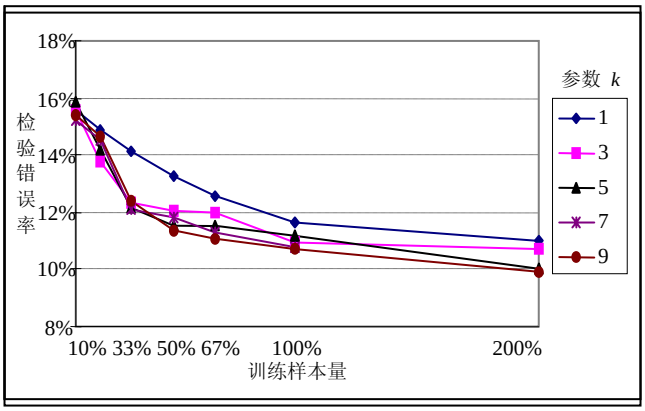

图 5-a: kNN 模型的样本量增量分析

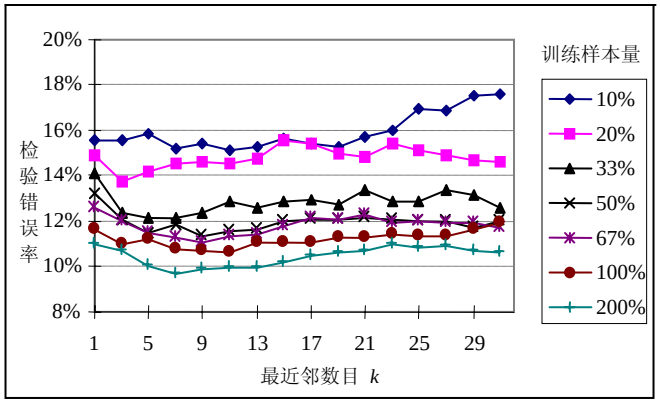

图 5-b: kNN 模型的复杂度增量分析

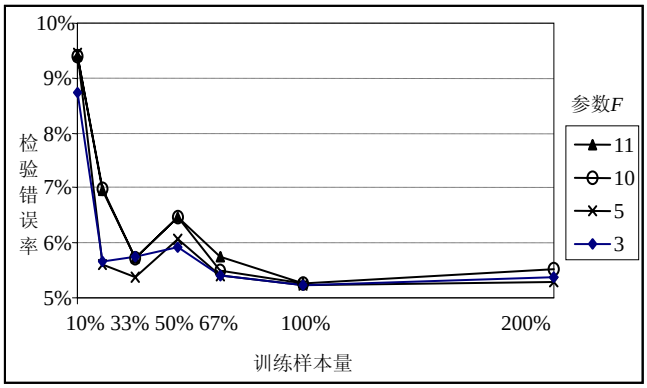

图 6-a: M5 模型的样本量增量分析

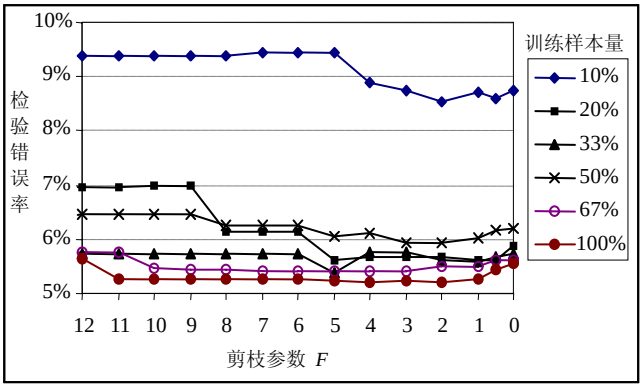

图 6-b: M5 模型的复杂度增量分析

回归树 M5 算法建立的决策树的复杂度由剪枝参数 *F* 决定，*F* 越大决策树的节点越少，

结构越简单。从总的趋势来看(见图6－a)，错误率不会因为增加训练样本量而获得进一步降低。因此，在后面建立 M5 模型时不增加训练样本。从图6－b看出，使用的样本数量越多，决策树的剪枝参数需要设定越大。对于100％的样本数，合适的模型复杂度为参数 $F = 11$。

4.2 第二阶段——比较分析

根据增量分析所确定的最优模型参数和训练样本数目，重新抽取训练样本建立最终模型，并运用检验样本检验模型，得出 ROC 曲线，计算最终模型的曲线下面积－AUC。五个模型的 ROC 曲线见图 7。可以看出最优模型为用 M5 算法建立的回归树模型。

5 模型的实际应用

将上面比较分析得出的最优模型 M5 应用于所有的样本数据，模型评分范围为 0－1，评分越大代表信用越好。结果表明评分模型成功地拟合了专家的判断：97％的 *A/B* 类企业的评分大于 0.7，100％的 *D* 类企业评分小于 0.01，*C* 类企业的评分分散于 0-0.94 之间。在实际应用中，可以通过灵活选取分割值的方式，决定 *C* 类企业信用状况。例如，可以选取两个分割值 *T1* 和 *T2*, *C* 类企业信用状况可以按照所得分值分成三种：“信用好”(评分大于 *T2*，申请的保险予以批准)、“未决定” (评分介于 *T1* 和 *T2* 之间，由专业人员进一步评判)和“信

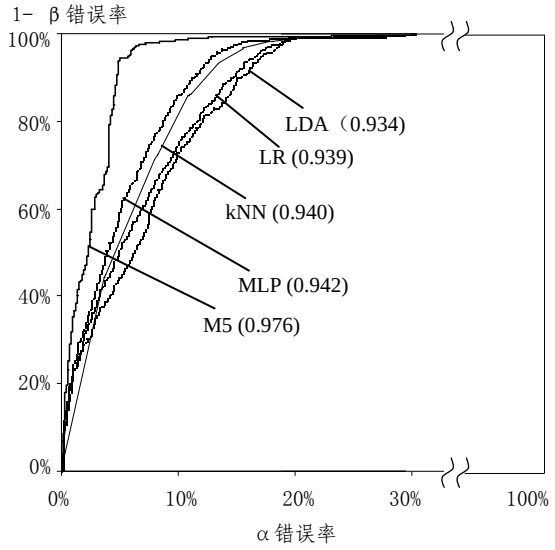

图 7.五种模型的 ROC 曲线及其 AUC(括号内)

用差” (评分小于*T1*，拒绝申请)。不同的分割值可以代表不同的信用政策——或宽松或谨慎，在实际应用中，可根据当时市场情况以及政策的倾向，灵活地调整所选取的分割值（见表3）。通过分割值的运用，可以将*C* 类企业进一步细分，发现其中信用良好的。如表3所示，如果取 *T1*=0.03，*T2*=0.7,则属于*C* 类企业的11%可以被判别为信用良好的，它们就是以往被忽视的可以给保险公司带来收入增加的利基市场。

6 结 论

本文提出了系统地优化信用评分模型的两阶段模式：（1）在模型建立过程中运用增量分析以确定模型参数和选定合适数量的训练样本，（2）在模型评价过程中运用ROC方法选择最优模型。将这一模式应用于实际信用保险公司评分模型的建立，利用优化的信用评分模型辅助保险申请的审批，在原有技术难以判断其信用状况的企业中识别出其中信用良好的企业，从而有

助于为该信用保险公司创建和发现利基市场。

表 3. 运用不同分割值评判 *C* 类企业的信用

（表中百分数为占所有 *C* 类样本数据的百分比）

分割值	信用差 拒绝申请	未决定	信用好 批准申请
T1=0.03, *T2*=0.7	70%	19%	**11%**
T1=0.02, *T2*=0.5	52%	32%	**16%**
T1=0.03, *T2*=0.2	70%	5%	**25%**

备注：①. ROC（receiver operating characteristic）原用于信号检测中描述噪声信道的命中率和虚报率的权衡关系。另外类似的方法如 Lorentz diagram 和 lift curves, 其基本原理与ROC曲线一致。

参考文献

[1]方洪全, 曾勇. 运用多元判别法评估企业信用风险的实例[J]. 预测，2004，23(4): 65-68.

[2]于立勇, 詹捷辉. 基于 Logistic 回归分析的违约概率预测研究[J]. 财经研究, 2004, 30(9):15-23.

[3] Henley WE and Hand DJ. Construction of a k-nearest-neighbor credit-scoring system[J]. IMA Journal of Mathematics Applied in Business and Industry,1997, 8: 305-321.

[4] Desai VS，Conway DG， Crook JN，and Overstreet GA. Credit scoring models in the credit-union environment using neural networks and genetic algorithms[J]. IMA Journal of Mathematics Applied in Business and Industry, 1997, 8: 323-346.

[5] Fritz S and Hosemann D. Restructuring the Credit Process: Behavior Scoring for Deutsche Bank's German Corporate[J]. International Journal of Intelligent Systems in Accounting, Finance & Management, 2000, 9: 9-21.

[6] Galindo J and Tamayo P. Credit Risk Assessment using Statistical and Machine Learning Basic Methodology and Risk Modeling Application[A]. Proceedings of Computational Economics' 97 Conference[C], 1997.

[7] 刘 ，罗慧. 上市公司财务危机预警分析——基于数据挖掘的研究[J]. 数理统计与管理, 2004, 23(3):51-56.

[8] 石庆焱，靳云汇. 多种个人信用评分模型在中国应用的比较研究[J]. 统计研究， 2004，6: 43-47.

[9] Joos P, Vanhoof K, Ooghe H, and Sierens N. Credit classification: a comparison of logit models and decision trees[A]. Application of machine learning and data mining in finance[C], 10th European Conference on Machine Learning, Workshop notes 24, TU Chemnitz, Germany, 1998: 59-70.

[10] Weiss SM and Indurkhya N. Predictive Data Mining: A Practical Guide[M]. Morgan Kaufmann Publishers, Inc., San Francisco, California, 1998.

[11] Nayak GN, Turvey CG. Credit risk assessment and the opportunity costs of loan misclassification[J]. Canadian Journal of Agricultural Economics, 1997, 45: 285-299.

[12] Hand DJ, Henley WE. Statistical Classification Methods in Consumer Credit Scoring: A Review[J]. Journal of the Royal Statistical Society, 1997,160, Part 3: 523-541.

[13] Hand DJ. Construction and Assessment of Classification Rules[M]. Wiley series in probability and statistics, John Wiley & Sons, 1997.

[14] Kelly MG, Hand DJ. Credit scoring with uncertain class definitions[J]. IMA Journal of Mathematics Applied in Business and Industry, 1999, 10: 331-345.

[15] Hand DJ, Adams NM. Defining attributes for scorecard construction in credit scoring[J]. Journal of Applied Statistics, 2000, 27(5): 527-540.

An empirical study on the optimization of credit scoring models and the discovery of a niche market for an insurance company

LIU Yang, LIU Weijiang

（Business School of Jilin University, Jilin Changchun 130012, China）

Abstract: The application of data mining classification techniques in credit scoring involves the process of building and selecting optimal models. The final model to be put into practice is decided through the efficient use of available samples, the determination of suitable model structure for various classification algorithms, and the evaluation and comparison of many alternative models. The paper presents an empirical study on this optimization process. The main results are: First, propose a two-stage optimization process for credit scoring models: (1) using the incremental analysis to determine optimal model complexity and suitable sample size for various algorithms, (2) then using ROC to evaluate and compare the accuracy of credit scoring models; second, apply the optimization process to a practical problem, the selected optimal model is used in discovering a niche market for a credit insurance company.

Keywords：credit scoring models；model evaluation; classification techniques; niche market

收稿日期：2006-2-10

基金项目：教育部留学回国人员科研启动基金项目（20041022），吉林省科技发展计划项目（20040304）。

作者简介：刘扬（1970-），吉林大学商学院副教授。