文章编号: 1000-6788(2009)12-0088-06

个人信用风险计量: 双边抗体人工免疫概率模型

杨 雨 [1], 史秀红 [2]

(1. 中央财经大学 管理科学与工程学院, 北京 100081; 2. 中央财经大学 金融学院, 北京 100081)

摘 要 研究了个人信用风险的计量问题, 构建了基于人工免疫机制的个人信用风险模型, 提出了双边抗体人工免疫概率模型, 利用商业银行实际数据进行了计算. 应用 ROC 方法对模型的预测能力进行了检验, 并与逻辑回归方法进行了比较, 达到并超过了逻辑回归模型的预测水平. 该模型系统不仅可以应用于个人信用的度量, 也可以应用于公司类客户的信用风险的度量以及电信和公共服务等领域.

关键词 双边抗体; 人工免疫; 概率模型; 违约概率

中图分类号 F830.73 **文献标志码** A

Personal credit risk measurement: Bilateral antibody artificial immune probability model

YANG Yu[1], SHI Xiu-hong[2]

(1. School of Management Science & Engineering, Central University of Finance and Economics, Beijing 100081, China;
2. School of Finance, Central University of Finance and Economics, Beijing 100081, China)

Abstract This thesis presents the approach to constructing consumer credit risk model by analysing personal credit risk for commercial banks. It also presents Bilateral Antibody Artificial Immune Probability Model and calculated the credit score using actual data from commercial bank. Test the forecast capability of model with ROC, and compare the result with that of logistic regression. This capability is sensitive to quantity of sample data, the more quantity are, the more forecast capability is. This model is well applied to forecast probability of default for consumer. This model can use other extensive domain.

Keywords bilateral antibody; artificial immunity; probability model; probability of default

1 引言

《巴塞尔新资本协议》提出信用风险内部评级法, 大幅度提高了资本监管的风险敏感度, 从而将资本充足率与银行信用风险结合起来 [1]. 作为消费信贷的主体, 商业银行在开展消费信贷业务的同时需要控制风险. 因此迫切需要一种系统化的理论方法, 为衡量个人客户的信用风险提供基准, 为商业银行建立全面风险管理体系提供技术支持, 为拓展个人信贷业务提供核心技能.

解决衡量个人信用风险的模型方法包括线性回归、非线性回归、分类树、判别分析等方法 [2]、专家系统、神经网络 [3]、遗传算法 [4−5] 等方法. 研究表明个人信用风险模型的发展趋势: 从线性化的方法到非线性化的

收稿日期: 2008-04-30
资助项目: 中财 121 人才工程青年博士发展基金 (QBG0701); 国家自然科学基金 (70773124); 中央财经大学 “211 工程” 三期
作者简介: 杨雨 (1969–), 博士, 讲师, 硕士生导师, 主要研究方向为风险管理, 风险计量.

方法; 从简单的结构化模型到非结构化的模型; 从参数模型到非参数模型; 从非智能的统计方法到优化技术再到人工智能的方法. 智能化的非线性非参数模型是信用度量的发展趋势.

应用免疫机制的原理研究商业银行个人信用风险问题, 以便在免疫系统框架下建立个人信用风险的系统化模型, 克服目前方法存在的某些缺陷. 本研究依据的原理和构造的模型系统不仅可以应用于对个人信用的度量, 也可以应用于公司类客户的信用风险的度量以及电信和公共服务等其他领域.

2 人工免疫原理与个人信用风险度量问题的相似性

2.1 个人信用风险模型与人工免疫原理的目标一致

个人信用风险模型与免疫系统的目的十分相似, 都是从众多个体中挑选少量不良个体.

个人信用风险模型的目标是根据个人的基本特性以及以往的行为特征预测未来个人客户发生违约的可能性以及违约造成的损失. 正常客户占全体客户的绝大部分, 违约客户只占全体客户的一小部分. 模型的核心是通过个人客户各种特征区分正常客户和违约客户, 并计算违约概率水平. 免疫系统中人体的 B 细胞的抗体识别外界病原体. 人体利用免疫屏障阻止外界的病原体侵入, 首先要利用系统中的抗体识别有害的抗原. 人工免疫系统计算抗体与抗原的亲和力水平, 当亲和力到达一定阀值时, 系统认定这种抗原属于有害抗原 [6].

2.2 个人信用风险模型与人工免疫原理的特征表达类似

两者的特征表达方式十分类似, 进入变量都是经过预先筛选, 表达客户整体风险特征.

个人信用风险模型实际上是通过群体特征预测个体客户的违约状况. 群体特征使用模型的参数表现出来. 进入模型的变量事先进行了筛选, 保证了这些进入变量既可以完整表达群体特征, 信息又不存在重叠. 人工免疫系统中通过学习后的抗体特征保存在记忆细胞的基因座中. 通过计算可以得出抗体中哪几个基因座可以表达抗原是否有害的特征 [7]. 抗体与抗原的亲和力实际上是基因的部分匹配.

2.3 个人信用风险模型与人工免疫原理的目标一致

两者的计算结果也非常类似, 都是得到 [0,1] 区间一系列数字. 人工免疫系统计算的亲和力与违约概率相可以找到映射.

个人信用风险模型计算每个申请者的违约概率, 即得到的违约概率用 [0,1] 区间一系列数值表示. 判断申请者优劣的标准是违约概率大小. 如果违约概率超过预先设定的阀值, 系统拒绝申请. 人工免疫系统计算进入系统中的抗原和抗体的亲和力, 也得到一系列 [0,1] 区间上的一系列数值. 这些数值表示进入抗原与已知抗体的接近程度.

2.4 个人信用风险模型与人工免疫原理的运行机制比较类似

判别个人信用风险的过程与人工免疫机制十分类似. 首先, 在判定新的贷款申请者前, 根据以往的申请者的表现的好坏按照他们的特征进行学习, 建立模型. 这相当于人工免疫机制的学习功能. 由此形成免疫系统的抗体特征; 当新的申请者申请贷款时, 需要利用模型进行判断. 这相当于人工免疫机制的模式识别机制. 也就是新的贷款申请者的特征相当于抗原特征, 抗原特征与抗体特征按照一定的规则进行匹配, 其亲和度函数的值就是表明新申请者的好坏; 新的贷款申请者进入后, 势必改变原有抗体的特征 [8]. 进化机制保证将新的特征添加到原有抗体特征中并进行记忆, 从而实现模型的进化.

3 双边抗体人工免疫概率模型

免疫系统由免疫学习、模式识别、免疫进化、免疫记忆等几部分组成, 实现机体的免疫应答 [9]. 相应地, 双边抗体人工免疫概率模型结构也相应包括学习机制、模式识别、进化机制和记忆机制这几个主要部分 (见图 1).

一般而言, 受到环境的影响、竞争的压力以及借款人的信用观念的变化, 应用个人信用风险模型几个月或者一年后都要进行模型参数的调整或者模型结构的优化. 但是利用人工免疫系统中的学习机制可以学习新的抗体特征, 当出现新的病原体形式, 利用记忆机制可以将新获取的特征保存在抗体特征中, 从而完成进化机制, 使抗体有指导地进化 [10]. 当符合这种抗体特征的抗原出现时, 抗体就能快速识别这种抗原. 这一点

也符合个人信用风险模型计算的快速性和系统的响应时间的要求. 应用人工免疫原理构建的个人信用风险模型体系具有人工智能的特点, 避免了结构化的模型定期需要进行参数调整和结构优化的困境 [11].

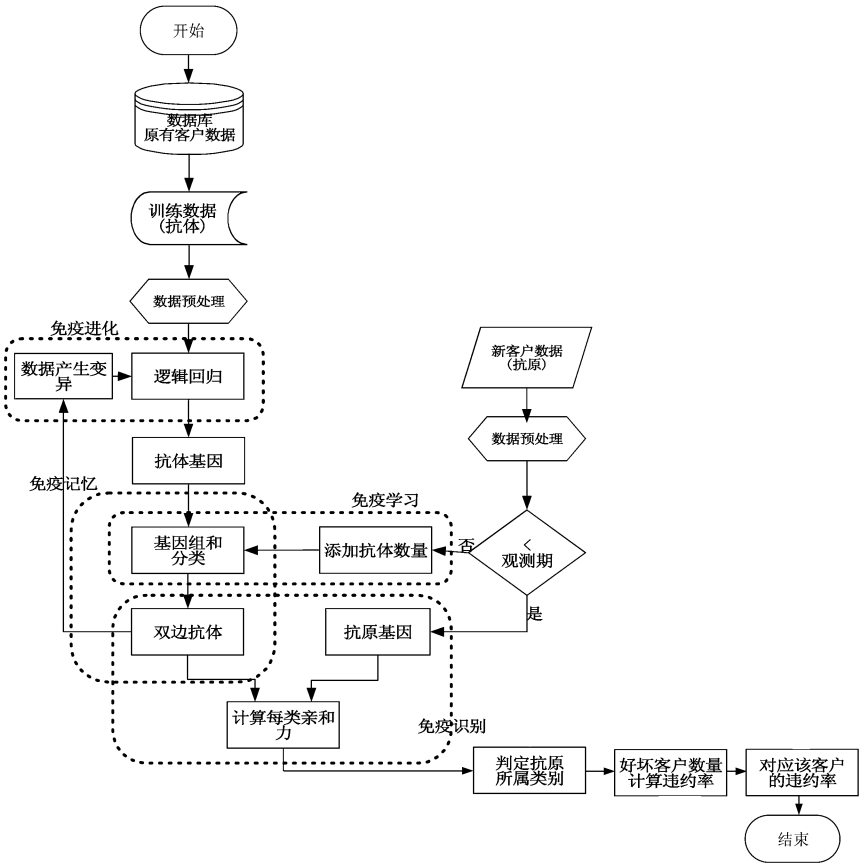

图 1 基于免疫机制的个人信用风险模型

4 双边抗体人工免疫概率模型的实证研究

4.1 数据来源

本研究的数据来源于某商业银行的个人贷款数据库. 从某商业银行的个人贷款系统数据库中随机抽取 2000 个样本作为研究样本. 这样样本数据和整体数据保持同一分布, 保证模型的预测能力, 避免预测模型过度拟合现象发生. 在 2000 个样本数据中, 实际只有 1991 个有效样本数据. 每个数据包括 92 个变量. 为了便于研究, 作数据清洗, 剔除那些带有缺失数据的样本. 剩余 1942 个数据样本, 分为三部分, 一部分作为训练样本、其余三部分作为学习样本一部分作为检验样本.

4.2 抗体的基因提取

由于个人客户与公司类客户不同, 个人客户的变量多为二分变量或者多值变量. 将指标进行规一化处理后, 采用 Logistic 逐步回归的方法进行回归计算, 回归结果就可以从中得到最利于好坏客户分类的指标. 抗体中的基因座保留原始的指标顺序, 客户特征发生变化后重新进行逻辑回归得到新的最利于分类的指标. 亲和力计算时只使用抗体基因座上最有解释力的那部分数据, 其余的仍然保留.

使用逻辑回归目的是得到分类预测能力最高的变量, 也就是抗体识别抗原时进行比较的基因座. 分别剔除那些对分类没有贡献的变量、不符合经济学含义的变量、不属于个人客户自然属性的变量. 对于筛选后的变量计算单变量的受试者作业曲线的值 (ROC), 按照降序进行排序. 使用 Logistic 回归得到变量对分类的贡献度最大的变量排序.

考虑业务因素, 剔除不必要的变量, 余下的 4 个变量重新使用 Logistic 逐步回归法, 可以得到下面的模型.

此模型的变量中职业、最高学历、岗位性质、居住状况的参数估计分别是 2.7809、2.6615、6.6426、9.6579, 而且显著性水平都在 97%以上. 因此该模型可以被用来预测个人客户的违约状况.

$$y = -10.942 + 2.7809 \times x_{28} + 2.6615 \times x_{27} + 6.6426 \times x_{33} + 9.6579 \times x_{52}$$

其中, 职业、最高学历、岗位性质、居住状况使用的是规一化后的值.

实际起作用的基因座如图 2 所示.

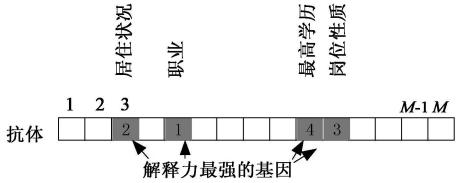

图 2　抗体中解释力最强的基因

4.3　免疫记忆

免疫系统中抗体浓度决定抗体识别的效率, 浓度越高, 抗体与抗原结合的可能性越大, 速度越快. 浓度升高通过抗体克隆完成. 由于记忆细胞数量有限, 所以免疫系统记录的抗体数量有限. 在个人风险问题中, 不良客户的信息被完全保留下来, 不良客户的数量也同时被保留. 系统记忆每类抗体的数量分别放入双边抗体的记忆单元中. 正常客户每类抗体的数量用 u_i 表示, 违约客户抗体的数量用 v_j 表示 (见图 3).

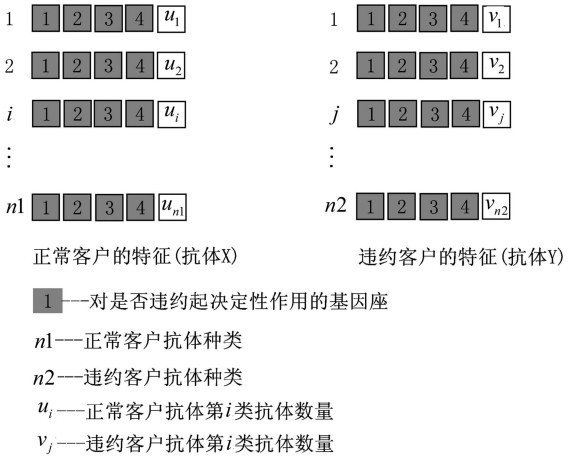

图 3　双边抗体记忆

4.4　免疫识别

免疫系统中抗体和抗原进行特异性识别, 实际上抗体群中与抗原最接近的抗体与抗原发生结合. 这种接近程度用亲和力表示. 按照解释力最强的基因座计算抗体和抗原的距离. 距离可以采取欧式距离.

$$\text{欧式距离: } r_i = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{ik})^2}$$

我们知道, 在这个样本空间分辨违约与否的基因的位置, 那么就可以建立亲和力与违约概率的关系. 计算出的亲和力只表明抗原与抗体的相似程度, 并不能直接转化为违约概率. 把初始抗体分为两组, 一组为正常客户的抗体集合, 另一组为违约客户的抗体集合, 形成这种双边抗体的结构. 免疫系统中的抗体都是单边的, 也就是免疫记忆细胞只记录入侵者的基因特征. 由于我们的问题不但是要鉴别个人客户是否违约, 还要求出他的违约概率. 单边抗体的结构由于无法建立参照系, 所以不能达到解题的要求.

$$\max(\min(RG_x), \min(RB_y))$$

首先计算新申请的客户也就是抗原与每一类抗体的距离, 分别用 RG 和 RB 表示. 分别取双边的最小值, 再取两个值中的最大值, 作为衡量抗原所属类别的范围.

$$Pd = \frac{\sum v_j}{\sum u_i + \sum v_j}$$

在此范围的类别中所有正常客户的抗体数量和非正常客户的抗体数量估计违约率.

抗体种类有限的另一个优势是计算快速. 虽然我们把所有训练样本和学习样本都进入抗体, 但是由于有效基因位置有限, 所以抗体空间只是基因座上基因所有取值的组合, 抗体的种类是非常有限的. 这样我们就可以达成在特异性识别和计算速度上的平衡.

4.5 免疫学习

进入系统的新客户超过观察期后, 违约状态就可以判定. 将这部分客户按照基因组合添加到系统的抗体中去, 改变每类组合中抗体的数量. 从而每类双边抗体的数量会随着时间的推移发生变化, 导致预测结果发生变化. 这一过程是一个自适应过程, 也是智能模型的重要环节.

4.6 免疫进化

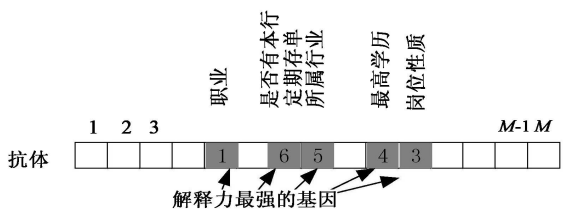

图 4 变异后的基因

利用免疫原理, 智能学习抗原特征. 并转化为新的抗体特征. 新数据的加入有可能使抗体特征发生变化. 于是形成了进化机制. 使用全部数据进行训练, 也就是学习了抗原特征的系统. 得到如下极大似然估计. 抗体的特征得到了进化. 我们来观察一下进化以后的抗体是否具有更强的识别能力 (识别能力靠 ROC 来测量). 此时识别是否违约的基因座发生了变化. 变量岗位性质消失, 添加了所属行业和是否有本行定期存单两个变量.

5 模型的检验

5.1 模型检验方法介绍

国际性的评级机构对个人信用风险模型的检验都采用 ROC 方法. ROC 曲线, 又称相对工作特征 (Relative operating characteristic). 任何阈值对应的模型表现都可以在 ROC 曲线图中得到体现 [12]. 国际上著名的信用评级公司都是使用 ROC 曲线和 ROC 的值来评估它们使用逻辑回归方法制作的信用风险模型的.

表 1 用于个人信用风险模型检验的 ROC 曲线的指标

诊断结果	实际结果		
	违约	不违约	合计
违约	a	b	$a+b$
不违约	c	d	$c+d$
合计	$a+c$	$b+d$	$a+b+c+d=N$

$$\text{灵敏度} = \text{真违约} = \frac{a}{a+c}$$

$$\text{特异度} = \text{真违约率} = \frac{d}{b+d}$$

ROC 曲线以假违约率为 X 轴, 以真违约率为 Y 轴. 该曲线可以通过以下步骤来构造. 以假不违约率为横轴, 以真违约率为纵轴, 横轴和纵轴长度相等, 形成正方形, 在图中将 ROC 工作点标出, 用直线连接相邻两点构建光滑 ROC 曲线.

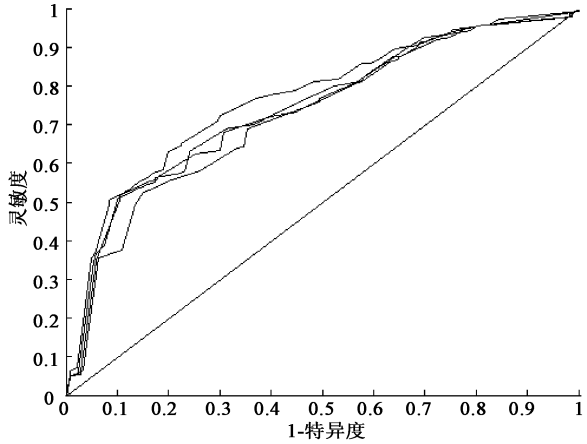

图 5 训练样本和添加三个抗体模型的 ROC 曲线

将 ROC 曲线进行 0 到 1 的积分, 则 ROC 曲线下的面积就是 ROC 值. ROC 的值越大代表模型的预测效果越好. 图中显示, 经过 3 次学习后模型的 ROC 曲线下的面积越来越大. 说明, 通过不断学习, 模型的预测能力得到不断提高.

5.2 模型间的预测能力的比较

双边抗体免疫概率模型的预测能力良好. 目前国际上主流商业银行采用逻辑回归的方法构建模型见 4.2 节, 所以在相同的数据条件下, 本研究与逻辑回归模型进行比较.

表 2　样本数量对模型 ROC 值的影响

模型样本数量					双边抗体免疫概率模型的 ROC 值				逻辑回归模型 ROC 值
训练	学 1	学 2	学 3	检验	训练	训练 + 学 1	训练 + 学 1+ 学 2	训练 + 学 1+ 学 2+ 学 3	训练
400	400	400	400	342	0.6160	0.7416	0.7530	0.7947	0.7874

表 2 所示, 当检验样本的数量一定时, 双抗体人工免疫概率模型的 ROC 值大于逻辑回归模型的 ROC 值.

逻辑回归模型是一种非智能化模型, 因此没有学习机制, 模型参数不能调整. 然而双边抗体人工免疫概率模型是一种智能化模型抗体需要不断添加. 从业务上逻辑回归模型建立以后一般要使用一段时间, 其间的参数不会变化, 因此两种模型在建模时使用的数据量不同, 双边抗体人工免疫概率模型使用全部数据效果更好, 而逻辑回归必须使用抽样数据.

6　结论

本研究提出了双边抗体人工免疫概率模型的系统框架, 实现了模型的免疫学习机制、免疫记忆机制、免疫识别机制和免疫进化机制. 通过银行实际数据进行了计算. 应用 ROC 方法对模型的预测能力进行了检验, 并和逻辑回归模型的预测能力进行了比较. 结果表明双边抗体人工免疫概率模型对个人信用风险的预测良好, 其检验值接近并超过了逻辑回归模型.

参考文献

[1] The New Basle Capital Accord(2nd Edition)[R]. 巴塞尔银行监管委员会, 2001.

[2] Barth J. A simple credit risk model with individual and collective components[R]. Mannheim, 1999.

[3] 冯铁军. 基于 GA 神经网络的个人信用评估 [J]. 上海金融, 2003(3): 48–51.
Feng T J. Indivadual credit evaluation based on genetic arithmetic neural networks[J]. Shanghai Finance, 2003(3): 48–51.

[4] 童烁, 费良俊. 发现金融市场预测模型的计算智能方法 [J]. 软件学报, 1999, 10(4): 395–399.
Tong F, Fei L J. Computational intelligence approach for discovering the prediction model of financial market[J]. Journal of Software, 1999, 10(4): 395–399.

[5] 袁礼海, 宋建社, 毕义明, 等. 混合遗传算法及与标准遗传算法对比研究 [J]. 计算机工程与应用, 2003(12): 124–125.
Yuan L H, Song J S, Bi Y M, et al. Research of contrast between hybrid genetic algorithm and genetic algorithm[J]. Computer Engineering and Applications, 2003(12): 124–125.

[6] Tarakanov A, Dasgupta D. A formal model of an artificial immune system[J]. BioSystems, 2000(55): 151–158.

[7] Timmis J, Neal N, Hunt J. An artificial immune system for data analysis[J]. BioSystems, 2000(55): 143–150.

[8] Hunt J E, Cooke D E. Learning using an artificial immune system[J]. Journal of Network and Computer Applications, 1996(19): 189–212.

[9] Gutnikov S, Melnikov Y. A simple non-linear model of immune response[J]. Chaos, Solitons and Fractals, 2003(16): 125–132.

[10] Timmis J, Neal M. A resource limited artificial system for data analysis[J]. Knowledge-Based System, 2001(14): 121–130.

[11] Yang Y, Liu Y. Study of consumer credit risk model based on artificial immunity mechanism[C]//International Conference on Innovation and Management, 2005.

[12] Hanley J A, et al. The meaning and use of area under a receiver operating characteristic curve(ROC)[J]. Radiogy, 1982, 143(1): 29.