密级：_____

人大经济论坛 www.pinggu.org
经 管 类 知 识 的 互 动 学 习 平 台

人大经济论坛评谷®数据处理和分析系列项目报告

神经网络模型在银企客户识别中的应用 (节选)

客户挖掘项目组

Tel: 86+010-684 542 76

Email: research@pinggu.org

人大经济论坛评谷®数据处理与分析中心

http://data.pinggu.org

1 客户识别的立项背景

银企的客户关系管理模式在进入互联网时代后发生了根本性的变化。一方面，由于金融产品和服务日益信息化和数字化，客户的期望也在迅速变化，产品同质化倾向也在增强。客户业务竞争愈加激烈，国有银企特有的竞争优势越来越难以获得。另一方面，计算机、通讯和网络技术的飞速发展，丰富多彩的数字信息已经在“点击鼠标的一瞬间”就可以获得。消费者的消费选择变得更富有信息和更加复杂化，能够根据所掌握的信息量来选择偏好的产品和服务。这就导致了银企和客户之间的互动关系正在发生着意义深远的变化。

1.1 CRM 新挑战

在拥有了广泛选择空间和极大选择便利前提下，银企的客户们完全可以决定要选择谁、何时选择和如何选择。互联网技术的迅速发展，使得这种选择局限已经摆脱了传统的空间地域关系，并且选择方式已经具备了“点击”瞬间的特征。客户对于随时随地得到服务的要求不仅越来越高，而且对质量、个性化和价值的要求更加挑剔。另一方面，对于银企来说，却面临着越来越多的客户关系管理（Customer Relationship Management ，CRM）方面的严峻挑战。

首先，客户识别方式的挑战。银企长期进行的规模营销方式受到了关系营销（一对一营销）方式的挑战。在传统营销过程中，营销目标是尽可能多地影响客户和扩大客户基础。但是，在市场竞争日趋激烈的情况下，由于获得新客户的成本比较高，与目前现有客户做交易相对要更好一些。所以，传统营销已经从客户基础的扩大转变为深化每一客户的需要。银企行为不仅是为了达成交易而与客户产生交往，而是将销售产品的机会转变为对客户服务的成功经历，并努力与每一有价值客户建立长期关系的具有时代特征的新型商务活动。这些必然产生了对于客户类型识别的新挑战。

接着，客户关系维持的挑战。银企面临着客户关系维持管理的挑战。信息多元化和网络的出现或者是存在，给银企造成了客户随时转移的潜在危险，同时也使客户在众多选择下由于不了解银企的有关诸如产品质量和价格成本等信息而使客户容易产生转移行为，即产生信息不对称下的客户逆向选择问题。同时对于银企来说，又会导致信息不对称下的道德风险问题。因此，客户的数据信息如何在多变的不对称信息条件下正确获得、客户关系如何在多变的不对称信息条件下正常维持、如何正确把握客户行为的数据信息来为银企决策服务，事实上已经成了银企当前在 CRM 领域中面临的非常重要的研究课题之一。

其次，产品或服务差异化的挑战。随着市场竞争日益加剧和客户需求的不断提升，银企面临着必须差异化他们的产品或服务的挑战。而应对这一挑战的一个有效方式，就是凭借一个能够准确地、连续地和客户接触的 CRM 管理系统，使银企自己和其它银企相区别。但是，要想实现该目标，银企必须要首先从收集客户人口和行为数据的基础工作开始，并建立银企本身的客户数据库和交易数据库，运用现代数据信息管理分析方法和先进的数据挖掘技术，对客户信息进行全面的分析，对客户进行准确评估，银企才能使精确的目标客户定位成为可能，根据客户对产品差异化和服务差异化的需求来确定经营战略。

从上述银企所面临的挑战来说，所涉及到的核心问题基本上可以归纳为两个方面，其一是来自于客户识别方面的挑战，其二则是来自于客户保持方面的挑战。从银企为了增加收益的角度来看，在应对这些挑战上，一般需要努力做到以下几个方面：即努力扩大银企的客户基础、努力增加有价值客户的数量同时减少无价值客户、高风险高成本客户和欺诈性客户的数量，以及努力提高客户对银企的忠诚。为此，银企需要力争做到：

1. 识别客户。完整掌握客户信息，分析客户购买行为和习惯，识别银企最有利可图的客户，针对不同的客户实施不同的营销组合策略。
2. 挖掘客户。准确把握客户行为特点和要求，快速响应个性化需求，提供便捷的购买渠道、良好的售后服务与经常性的客户关怀；
3. 维持客户。将有关已经流失客户的历史数据与维系客户的控制群体进行比较，识别流失客户与众不同的行为，识别将要流失客户的行为模式，采取有效的防御性营销策略。

在这三个方面，识别客户是基础、挖掘客户是关键，而维持客户则是必要前提。所以，当前银企所面临问题的关键就在于如何正确识别客户和有效的做到客户挖掘。利用数据挖掘技术和客户数据库分析技术来考察客户行为，对于认清客户关系管理的本质，探索银企客户关系管理机制，帮助高层管理者根据事实做出以客户为中心的决策从而获得真正的竞争优势以及管理未来和提高增加收入的能力，具有十分重要的理论价值和现实意义。特别是在市场竞争日趋激烈而国内有关客户识别和客户保持问题的理论研究尚处在起步阶段的情况下，加强对这一问题的深入探讨就显得更加迫切。本报告也正是基于这样一种基本的认识，拟从银企的客户识别与客户保持的理论分析入手，建立客户识别与保持模型，以期有益于 CRM 的理论建设和实践管理上的进一步拓展和应用。

随着各银企分支机构不断的增多，行业竞争日益激烈，金融市场的供求格局目前已经发生了根本性转变，买方金融市场的特征已初步形成。国内金融市场已被国有银企及中小银行初步分割完毕，规模效益不再突出，资产质量日益成为银行的生命线。这使得各银企的经营策略都逐步朝着抢占垄断行业和优质客户的方向倾斜，客户成为银行至关重要的商业资源。但是由于行业垄断政策，银行长期处于保护状态，缺乏活力，业务流程还是基于内部管理和内部核算的需要，并没有把“以客户为中心”真正落到实处。如何实施管理客户资产、如何来科学地强化客户资产管理，无论是在经验上还是在技术上，都显得十分的缺乏。特别是在银企客户关系管理上，由于经验和技术的缺乏，银企还不能够从浩如烟海的海量客户数据中科学地识别客户和对客户市场进行有效的分割。

在加入 WTO 以后，银行业又面临着同欧美等外资金融机构竞争的局面，这类机构在诸多方面存在着显著的优势。外资银行的经营规模庞大，资金实力雄厚，资产质量优良，特别是在客户关系管理方面，国外已有多年的经验。而国有银行对如何通过客户关系管理向顾客提供真正的个性化服务，至今还没有找出一条好的办法。对“以客户为中心”的理解一直处于表面状态，不能够深入的了解客户的需求，长期以来对客户实行无差别服务策略，无论是老客户还是新客户，大客户还是小客户，都一律平等对待，不能够针对不同客户提供不同服务，不能够抓住真正的赢利客户并进行区别对待，为客户提供一对一的服务。此外，信息技术的飞速发展深刻地影响了银企的运作模式，网络经济与电子商务初现端倪，这就要求银企必须提供更加有效的金融服务。这些问题都是原有的经营战略、机制和技术所无法解决的，银行业迫切需要能帮助其全面联系外部客户、把握市场、创造收益的 CRM。

1.2 网络模型意义

较早使用神经网络解决经济问题的为White (1988)，他使用了神经网络模型对 IBM 的公司股票收益进行了预测。Kuan and White (1994). 基于严格的数学推导，证明了神经网络技术和计量经济学方法之间的关系，认为神经网络是可以对经济理论以及计量经济学做出大贡献的。神经网络在经济领域的应用越来越广泛，这也验证了神经网络技术在建摸以及估计动态预测方面起到重要作用。Kuan and Liu (1995) 基于前馈式及复现式神经网络研究了汇率预测的问题，Haefke and Helmenstein (1996) 使用线性

模型与神经网络模型预测了奥地利股市的 IPO 问题，Lisi and Schiavo (1999) 则对比了神经网络模型与混沌模型对于汇率的预测效果。进一步，Kodogiannis and Lolis (2002) 基于 MLP 神经网络、RBF 神经网络、动态神经网络以及模糊神经网络模型预测了美元与英镑的日汇率变化。此外，Hagan et al. (2002) 所著《神经网络设计》一书，以深入浅出的方式诠释了基本的神经网络模型结构以及学习算法，被誉为神经网络模型的经典教科书。

对于银企而言，成功地获得有价值的新客户和维系高价值客户是至关重要的。它们通常为此建立了大型数据库，可以分析和应用于开发新的商务策略和机会。然而，不是同等地瞄准所有的客户或对所有客户都提供同样的激励，银企可以仅选择那些个人需要和购买行为符合一定利润标准的客户 (Dyché, 2002)。关于信用、行为的客户识别模型就是帮助银企决策制定者了解和识别个体客户的方法。这类模型根据客户的特征诸如年龄、收入和有关的状况帮助银行确定是否将信用给予新申请人（如是否贷款给某个客户），帮助银企分析现有客户的购买行为 (Setiono et al., 1998)。客户识别模型是统计和运作研究模型在金融和银行行业最成功的应用，并且越来越多的企业和研究者采用它们来分析和识别客户。Chen and Huang (2003) 使用信用分数的分类模型根据年龄、收入等特征对新贷款申请者进行分类；Thomas (2000) 使用行为分数分类模型，根据信用分数变量和其它的描述行为的变量预测现有客户未来的信用状态；Malhotra and Malhotra (2003) 使用神经网络构建分数模型以分析和管理个体客户贷款；Kim and Sohn (2004) 运用神经网络方法使用信用分数模型的错分类模式管理贷款客户。

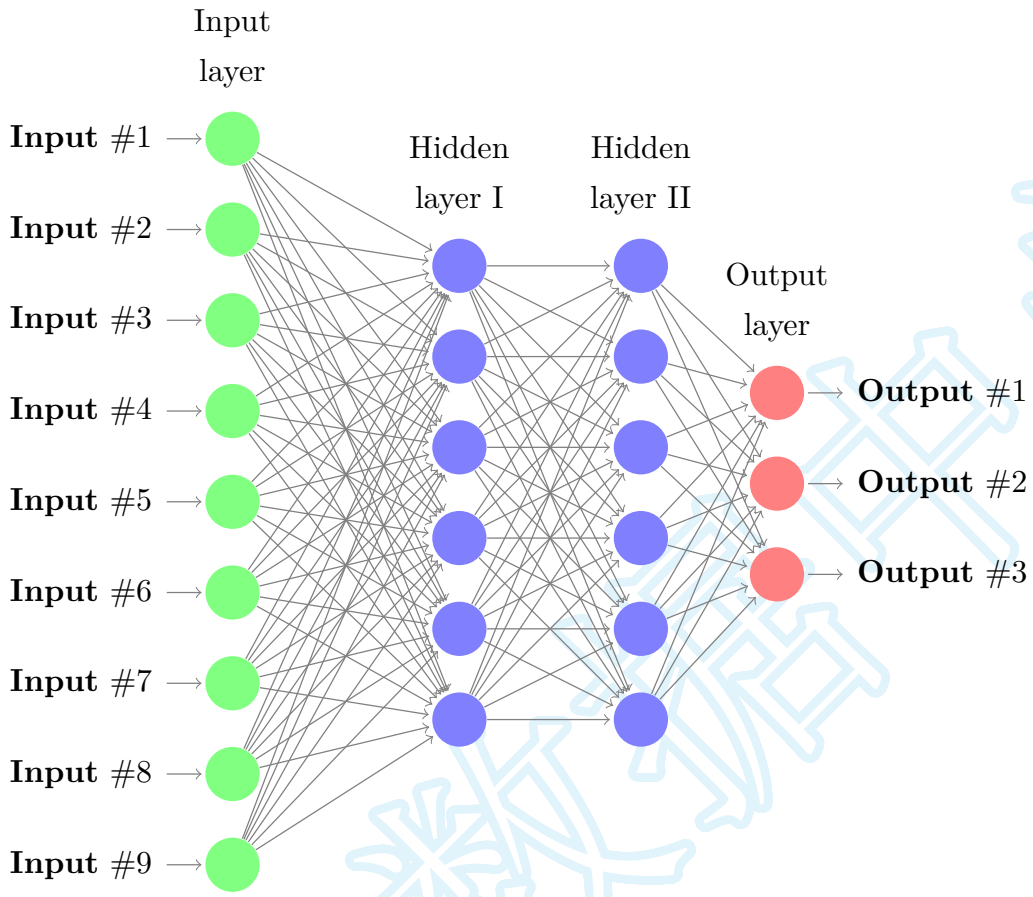
银企的客户识别及挖掘是一项长期而复杂的任务。银企客户的价值及其管理。信息时代银企面临的客户关系管理的实质是客户价值的管理，如何建立客户价值管理的框架，识别客户识别及保持客户是客户价值管理的主要内容。经济全球化以及电子商务的实施，使银企面临更加激烈的竞争，有效管理客户资产的能力已成为银企生存和发展的决定性因素。本项目对银企客户识别和客户保持问题的研究对于认清客户关系管理的本质，探索银企客户资产管理的运作机制，提高银企基于事实的决策制定能力，实现低成本、高质量、快速响应市场需求和客户需求的目标，有效率地分配银企稀少的资源，改善与客户的关系，帮助银企获得真正的竞争优势和增强银企管理未来和增加收入的能力，具有十分重要的理论价值和现实意义。本项目正是在这样的背景下，运用神经网络模型（Neural Network Model）的技术，围绕具有广泛的实际背景和发展前景的银企客户关系管理决策与优化问题，对银企客户识别和客户保持的决策模型进行了深入研究。

客户关系考察指标之间的关系通常是复杂，基于高度简化的线性模型来刻划这类关系是困难而不切实际的。然而，使用复杂非线性模型的，又会削弱经济模型的解释能力。这就需要在模型的复杂性与经济的可解释性之间寻求一种均衡。神经网络模型便为这种均衡提供了现实的可能性。这类模型突出优点为：一方面能够以线性与非线性函数的组合形式很好地逼近自变量与因变量之间的几乎任意函数关系，另一方面又能够通过优化隐藏层神经元数量、修剪神经网络链接及单元的方式来简化模型。神经网络模型是用来模拟人脑结构及智能特点的一个前沿研究领域，它的一个重要特点是通过网络学习达到其输出与期望输出相符的结果，具有很强的自学习、自适应、稳健性、容错性及存储记忆的能力。

神经网络模型的优点。神经网络由许多神经元共同组成，总的来说，大致都有以下特征：

1. 非线性拟合。神经网络有近似任意非线性能力，这使其具有非线性变换特性，为以后线性问题的解决带来新的希望。
2. 并行处理。神经网络具有紧密的并列结构与并列操作功能，所以具有很好的耐故障能力与快速处理能力。对于既时控制与动态控制两种控制非常合适。

图 1: 双层感知器的神经网络模型示例



注意：在引入了隐藏层神经元的条件下，神经网络模型能够以优化方式进行分类与记忆，从而该模型的学习算法成为了人工智能研究的热点。

3. 集成效应。神经网络可以在线操作，具有定量和定性同时操作的功能。神经网络可以同时输入许多不一样的信号，解决信息之间的问题，适用于控制复杂、多变量和大规模系统。
4. 训练学习。神经网络运用对过去数据记录进行分析而训练的。经过训练之后的神经网络能够对数据进行归纳。所以神经网络可以解决数学模型或者描述规划无法解决的问题。

神经网络系统评价方法以其超凡的处理复杂非线性问题的能力独树一帜，这种评价方法忠实于客观实际，不带任何人为干预的成分，是一种较好的动态评价方法。近年来，神经网络模型的研究和应用受到了的极大重视。神经网络模型具有多种类型，应用最为广泛的是多层感知器（MLP）模型和径向基函数（RBNF）模型。。MLP 模型是一种具有单层或多层的阶层型神经网络，层次之间的各神经元实现完全连接，即前一层的各神经元与后一层的各神经元都实现权连接，而各层之内的神经元之间没有连接。MLP 模型通常在隐藏层使用线性组合函数和 S 型激活函数，在输出层也使用线性组合函数但激活函数在需要与前者相适应图1显示了双层感知器的 MLP 模型结构。本项目应用 MLP 神经网络模型对银企的客户识别过程进行模拟和预测。

2 客户挖掘的模型构建

在客户价值分析和评估过程中，客户识别与挖掘的关键在于利用各种数据挖掘技术和统计技术建立模型，然后利用模型发现和定量描述模式，并对模型发现客户知识的能力进行评估，目的是通过银企过去的客户数据和交易数据获得客户知识并利用这些知识了解和预测未来客户的行为，以帮助银企获得管理未来和增加收入的能力。

2.1 理论基础

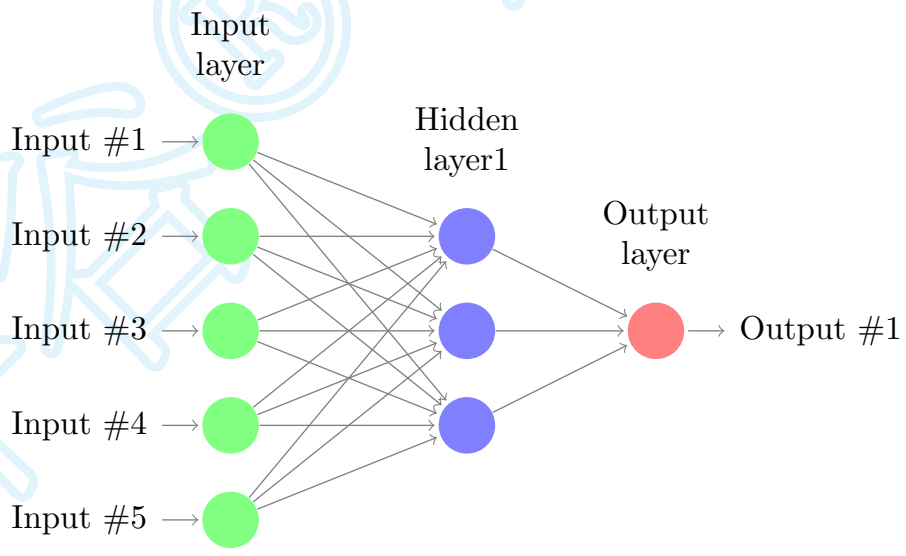
CRM 的目标是要借助于先进的数据仓库技术和数据挖掘技术，分析现有客户和潜在客户相关的需求、模式、机会、风险和成本，从而最大限度的赢得银企整体经济效益。这就要求运用数据挖掘技术从银企的客户数据库或数据仓库中挖掘出隐含的、事先未知的相关知识，识别优质客户，屏蔽欺诈客户，保持原有客户并吸引更多的潜在客户，提高经济效益和行业竞争力。

银企实施客户关系管理的基本原则是：明确客户收益点，增加盈利。这就需要在识别优质客户、屏蔽欺诈客户的基础上，维持旧客户并且增加新客户。进一步，需要运用数据挖掘技术从银企的客户数据库或数据仓库中挖掘出隐含的、事先未知的相关客户知识。考虑银企特定的客户知识概念，可以 CRM 过程以模型表示为

$$y_i = f(W, X_i) \qquad i = 1, 2, \cdots, n \tag{2.1}$$

其中，$f(\cdot)$ 为非线性函数，y_i 为客户类型，$X_i = \{x_{i,1}, \cdots, x_{i,m}\}$ 为客户价值属性的相关指标，$W = \{w_1, \cdots, x_n, \cdots\}$ 为权重序列。非线性函数 $f(\cdot)$ 为刻划客户属性与类型之间的关系是个极其复杂的过程，基于一般意义上的统计模型是难以实现的。然而，神经网络模型方法可以对非线性模型充分逼近。因此，使用这类模型来实现银企客户挖掘过程在理论意义上是完全可行的。

图 2: 单隐藏层的神经网络模型



注意：单隐藏层的神经元 3 个，输入指标为 5 个、输出指标 1 个。

图2显示了具有单层感知器的客户识别网络结构。假定在隐藏层使用线性组合函数和 S 型激活函数，在输出层使用线性组合函数与因变量相适应的激活函数。那么，可以将隐藏层或者输出层的设计为线性组合的函数形式

$$u = \sum_j (\cdot) = b_j + \sum_{j=1}^{s} w_{rj} v_r \tag{2.2}$$

进一步，可以选择激活函数为如下常用的 S 型函数

$$\text{Gauss} : A(u) = \exp\left(-u^2\right) \in (0,1] \tag{2.3}$$

$$\text{Logistic} : A(u) = [1 + \exp(-u)]^{-1} \in (0,1) \tag{2.4}$$

$$\text{Tanh} : A(u) = 1 - 2[1 + \exp(2u)]^{-1} \in (-1,1) \tag{2.5}$$

$$\text{Elliott} : A(u) = u[1 + |u|]^{-1} \in (-1,1) \tag{2.6}$$

$$\text{Arc tan} : A(u) = \frac{2}{\pi} \arctan(u) \in (-1,1) \tag{2.7}$$

$$\text{Softmax} : A\left(u_j\right) = \exp\left(u_j\right)\left[\sum_{j=1}^{J} \exp\left(u_j\right)\right]^{-1} \in (-1,1) \tag{2.8}$$

这样，即便是单层感知器也可以形成复杂的非线性模型。比如，如果隐藏层使用线性组合函数和 Logistic 激活函数，则三个隐藏神经元的输出为

$$h_1 = [1 + \exp(-b_1 - w_{11}x_1 - \cdots - w_{51}x_5)]^{-1} \tag{2.9}$$

$$h_2 = [1 + \exp(-b_2 - w_{12}x_1 - \cdots - w_{52}x_5)]^{-1} \tag{2.10}$$

$$h_3 = [1 + \exp(-b_3 - w_{13}x_1 - \cdots - w_{53}x_5)]^{-1} \tag{2.11}$$

如果输出层也使用线性组合函数和 Arctan 激活函数，则整个网络的输出为

$$\begin{aligned} \mu &= \frac{2}{\pi} \arctan(\alpha + \beta_1 h_1 + \beta_2 h_2 + \beta_3 h_3) \\ &= \frac{2}{\pi} \arctan\left\{\alpha + \sum_{i=1}^{3} \beta_1 [1 + \exp(-b_1 - w_{11}x_1 - \cdots - w_{51}x_5)]^{-1}\right\} \end{aligned} \tag{2.12}$$

多层感知器是 Universal Approximator。只要给予足够的数据、隐藏神经元和训练时间，含有一层隐藏层的多层感知器就能够以任意精度逼近自变量与因变量之间差不多任何形式的函数；更多的隐藏层则可能减少隐藏神经元和参数的数目，提高模型的可推广性。神经网络模型的结构具有很大的灵活性。每一层的各神经元并非一定要全部连接到下一层的各神经元，可以去掉一些连接；输入神经元也并非一定要经过隐藏神经元再连接到输出神经元，可以跳过隐藏层直接连接到输出神经元。

典型的神经网络模型可以视为广义线性模型的推广。令 μ 表示因变量 Y 分布的位置参数。广义线性模型中的系统成分使用连接函数 $\eta = g(\mu)$，再令 $\eta = \alpha + x'\beta$。在神经网络模型中，如果在输出层使用线性组合函数，再令 η' 为组合后的值

$$\eta' = \alpha + h'\beta$$

其中， h 为隐藏层各神经元的输出值组合的向量；设输出层的激活函数为 A，并令神经网络的输出值为 $\mu = A(\eta')$。如果让等于的逆函数，则 $\eta' = A^{-1}(\mu) = g(\mu) = \eta$。此时，神经网络模型相当于与广义线性模型使用了同样的连接函数。

设数据集为 $\{(x_i, y_i), i = 1, \ldots, N\}$， μ_i 为与观察 i 对应的 μ 值。根据 μ_i 与 y_i 的差距，可以定义误差函数。误差函数越小，模型拟合效果越好。

2.2 网络学习

神经网络的学习是通过神经网络所在环境的刺激对自由参数进行调整，让神经网络对外部环境改变做出反应的过程。这种模拟人脑活动规律的人工神经网络系统，在银企的客户识别和挖掘日常工作中能够发挥着关键性的重要作用。

神经网络的学习算法在网络收敛特性、学习速度、泛化能力等方面起很大的作用。神经网络模型近似函数以及信息处理的能力，基本由网络中各神经元的耦合权重所决定的。对于规模大的网络，权重不能完全进行统一设定。所以网络自身应有学习的功能，就是可以在示范模式的学习过程中慢慢进行权重调整，网络整体应具备近似函数以及信息处理的能力。总之，神经网络的学习最终还是权重的调整。确定神经网络权重包含形式为两类：一类为按照设计运算来确定，另一类则为通过学习训练得到。通常情况下，神经网络模型的学习过程采用了后一类方式。

神经网络的学习也可叫作训练，就是通过神经网络所在环境的刺激对自由参数进行调整，让神经网络对外部环境改变做出反应的过程。这种模拟人脑活动神经网络系统，其学习训练方法可以概括为以下几种。从学习过程的组织与管理来看，分为监督学习与无监督学习; 从学习过程的推理及决策方式来看，分为确定型学习、随机学习与模糊学习。下面对这几种学习算法的含义和特点进行简单的介绍:

1. 有监督学习。就有监督学习而言，网络训练通常是以一定数量的训练样例或样本，训练样本一般组成含输入矢量以及目标矢量。通过这种学习得出结果，按网络输出的评价标准与实际输出和的评价标准进行对比，依据对比结果或误差，按一定比例调节权重与闽值，令网络输出值向目标值靠拢。网络评价标准是通过外界提供的，在这种学习中常需要按照某种规则调整网络连接权重。
2. 无监督学习。这是一种完全自身进行组织学习，没有外部教师的示范，也没有外部环境作用的反馈，也没有环境的反馈指导其需要输出什么、输出的正确与否。在整个过程中，网络仅对输入信号的激励做出反应，自动调节网络权重和闽值勤，直至最后达到一种理想的有序状态。关于自组织学习，一般学习规则有 Hebb 学习规则，近学习规则，产生竞争效果。所以，自组织学习的时候，协调等来实现。产生放大效果; 还有相是利用自放大、竞争和关于无监督学习，其能够实现主分量分析、编码、聚类和特征映射等作用。
3. 混合监督方式。有监督学习虽然分类细密、准确，可是学习很慢。而无监督学习含有分类灵活和算法简练的优势，但精确性较低。所以令两者组合一起就能避免不足，发挥各自的优势，形成一种更加有效的学习方法。混合学习过程通常为利用无监督学习抽取输入数据的特性，之后把这内部信号转送到有监督学习部分来分析，从而实现输入与输出之间的对应关系。因为在数据输入之前就进行了数据预处理，这就加快了有监督学习乃至整个学习的速 2 神经网络与遗传规划度。

考虑到了银企的业务实际，本项目采用了基于误差信号不断验证的有监督学习方式。图2显示了 CRM 项目中的神经网络模型的训练学习过程。从图中可以看到，神经网络模型的学习过程是客户数据挖掘的关键部分，也可以说是决定了 CRM 策略成功的关键，只有当客户类型能够得到判断，这一策略才是有利可图的，并且成功的客户保持能够降低银行寻求新的、具有潜在风险客户的需求，并使银行将注意力集中在建立关系和满足现有客户的需求上。

2.3 模型设计

一般而言，神经网络模型的设计内容主要包含了系统的输入、输出设计与网络的设计，而网络主体的设计又包含网络拓扑结构的确定及其权重的学习。系统输入的设计包含训练集的选择、编码的设计和输入方式的确定。

1. 选择训练集。在传统的系统设计过程中，知识的组合与表示是一困难的问题，如何选取这杂乱无章的知识还尚未有一般的原则将它们组建起来。就神经网络的设计来说，假如将随意的知识传达到网络学习，也许会由于知识量大的网络系统，势必造成计算机资源的浪费。也有可能是由于知识覆盖面小而无法运作正常。因此，在神经网络设计的过程中，不必将所有信息处理单位的特征综合。做到具体问题具体分析，有计划的选取样本，尽力反映训练样本的特性。
2. 设计编码。在训练集确定以后，接下去是对训练训练集中的样本进行输入/输出编码，即让网络用何种形式来学习。设计编码同样属于知识的表达。编码的设计主要运用实验和比较的方法。神经网络的判断能比推理能力强。所以在设计编码时应尽量让其进行判断性的操作而并非进行推理。
3. 设计输入方式。人们在知识学习的过程会采取一定的次序与方法，例如进行类比或者对照，从难至易，还是从易至难。在神经网络的训练，也需要按照合理的方式设计输入顺序，从而提高学习效率，合理地组织知识，更好的解决问题。

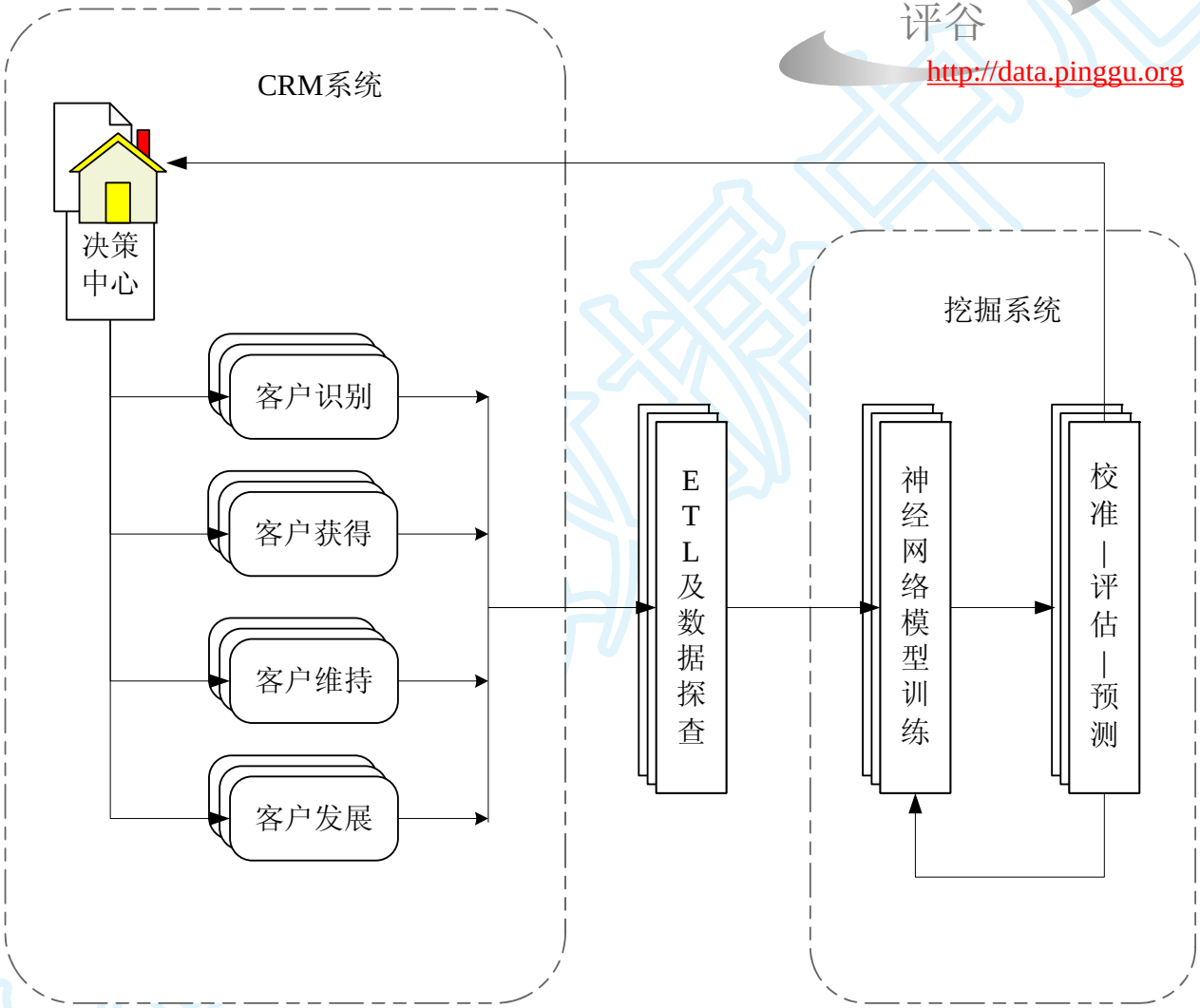
系统输出的设计在一个应用系统的设计过程中，难以将所有的问题避免，所以为了使系统更加可靠，可以独立列出 n 个子系统，这 n 个子系统能够按照不同的训练集和编码方案，之后对系统的输出处理，获得整个系统的输出。

神经网络的主体设计包含选择神经网络模型和确定参数。

1. 选择模型。经典的神经网络模型含几十种，但学习算法各异。这些算法的目的都是为了解决不同问题的需要，且都有优点和缺点。如果实际的网络系统完善，则可根据问题需要设计出相应形式的模型。网络模型的设计还应该与相应的学习算法保持一致性。模型不一样则合适的学习算法也不一样。比如，分类问题适于有监督的学习算法，聚类问题则适于无监督的算法。
2. 确定参数。当前应用最广泛的一种是多层前向网络神经网络模型。其主要参数包括隐层数、隐层结点数、连接权系数、传递函数、神经元的互连方式和学习规则等。神经网络的设计其实就是这些参数的确定过程。

设计神经网络的结构与参数，一般办法都是按照经验确定，可是在运用神经网络系统面临实际的问题的情况下，其解决问题的能力由说选择的网络结构与参数所决定，只是根据经验是难以寻求到适合的网络结构与参数，因此在近些年来研究神经网络的自动设计一直非常热门。

图 3: CRM 项目的模型训练学习



注意：CRM 项目当前只包含了客户识别部分。

3 数据整理及指标设计

数据整理时进行企业数据挖掘的前提和基础，占用了整个项目的大部分时间，其设计优劣直接关接到整个企业项目的成败。ETL 的实现有多种方法。常用工具如 Oracle 的 OWB、SQL server 2000 的 DTS、SQL Server2005 的 SSIS 服务、informatic 等实现，再者就是 SQL 方式实现，此外是 ETL 工具和 SQL 相结合方式。项目组考虑到进一步数据挖掘的需要，采用了 SAS/SQL 工具来进行银企的 ETL 过程。

3.1 ETL 过程

ETL 过程是将数据库系统的数据经过抽取、清洗转换之后加载进入数据仓库的过程。实现该过程的目的是将银企中的分散、零乱、标准不统一的数据整合到一起，为数据分析与挖掘建立必要前提。ETL 是商业智能最重要的一个环节，占用了整个项目的大部分时间，其设计优劣直接关接到商业智能项目的成败。ETL 的实现有多种方法。常用工具如 Oracle 的 OWB、SQL server 2000 的 DTS、SQL Server2005 的 SSIS 服务、informatic 等实现，再者就是 SQL 方式实现，此外是 ETL 工具和 SQL 相结合方式。项目组考虑到进一步数据挖掘的需要，采用了 SAS/SQL 工具来进行银企的 ETL 过程。

数据的抽取需要在调研阶段做大量工作。需要解决诸多问题：从不同类型的业务系统中发现数据；整合各业务系统的数据库服务器 DBMS；高效整理非结构化的数据以及手工录入部分数据等等。在收集完这些信息之后，才可以进行如下内容的数据抽取设计：

1. 同质数据源处理。数据源与 DW 数据库系统一致的条件下，数据处理工作会比较容易的。一般情况下，DBMS(包括 SQLServer，Oracle) 都会提供数据库链接功能，在 DW 数据库服务器和原业务系统之间建立直接的链接关系就可以写 Select 语句直接访问。
2. 异质数据源处理。在与 DW 数据库系统一致的条件下，数据源处理也可以通过 ODBC 的方式建立数据库链接，如 SQL Server 和 Oracle 之间。如果不能建立数据库链接，可以有两种方式完成，一种是通过工具将源数据导出成.txt 或者是.xls 文件，然后再将这些源系统文件导入到 ODS 中。另外一种方法通过程序接口来完成。
3. PC 文件 (.csv、.dat、.txt、.xls) 数据源的处理。可以培训业务人员利用数据库工具将这些数据导入到指定的数据库，然后从指定的数据库抽取。或者可以借助工具实现，如 SQL SERVER 2005 的 SSIS 服务的平面数据源和平面目标等组件导入 ODS（Operational Data Store）中去。
4. 增量更新。对海量的数据库系统，必须考虑增量抽取的做法。一般情况，业务系统会记录业务发生的时间，可以用作增量的标志。在每次抽取之前，判断 ODS 中记录最大的时间，再根据这个时间去业务系统取该时间后的所有记录。此外，可以考虑利用业务系统的时间标识，但一些系统没有时间标识。

数据清洗是发现并纠正数据文件中可识别的错误的最后一道程序。该过程包括检查数据一致性、处理无效值和缺失值等，任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。不符合要求的数据主要是有不完整的数据、错误的数据和重复的数据三大类。

1. 缺失数据。其特征是是一些应该有的信息缺失，如供应商的名称，分公司的名称，客户的区域信息缺失、业务系统中主表与明细表不能匹配等。需要将这一类数据过滤出来，按缺失的内容分别写入

不同 Excel 文件向客户提交，要求在规定的时间内补全。补全后才写入数据仓库。

2. 错误数据。产生原因是业务系统不够健全，在接收输入后没有进行判断直接写入后台数据库造成的，比如数值数据输成全角数字字符、字符串数据后面有一个回车、日期格式不正确、日期越界等。这一类数据也要分类，对于类似于全角字符、数据前后有不面见字符的问题只能写 SQL 的方式找出来，然后要求客户在业务系统修正之后抽取；日期格式不正确的或者是日期越界的这一类错误会导致 ETL 运行失败，这一类错误需要去业务系统数据库用 SQL 的方式挑出来，交给业务主管部门要求限期修正，修正之后再抽取。
3. 重复数据。这类错误在维度表中比较常见。必须将重复的数据的记录所有字段导出来，让客户确认并整理。

数据清洗是一个反复的过程，需要反复检查、发现及解决问题。数据仓库分为 ODS、DW 两部分，通常的做法是从业务系统到 ODS 做清洗，将脏数据和不完整数据过滤掉，再从 ODS 到 DW 的过程中转换，进行一些业务规则的计算和聚合。对于是否过滤、是否修正一般要求客户确认；对于过滤掉的数据，写入 Excel 文件或者将过滤数据写入数据表，在 ETL 开发的初期可以每天向业务单位发送过滤数据的邮件，促使他们尽快的修正错误，同时也可以作为将来验证数据的依据。数据清洗需要注意的是不要将有用的数据过滤掉了，对于每个过滤规则认真进行验证，并要用户确认才行。

数据转换的任务主要是进行数据一致性的实现、数据粒度的转换和业务规范的统一。

1. 一致性实现。这个过程是一个整合的过程，将不同业务系统的相同类型的数据统一，比如同一个供应商在结算系统的编码是 A0001，而在 CRM 中编码是 S0001，这样在抽取过来之后统一转换成一个编码。
2. 粒度转换。业务系统一般存储非常明细的数据，而数据仓库中的数据是用来分析的，不需要非常明细的数据，一般情况下，会将业务系统数据按照数据仓库粒度进行聚合。
3. 规范统一。不同的银企有不同的业务规则，不同的数据指标，这些指标有的时候不是简单的加加减减就能完成，这个时候需要在 ETL 中将这些数据指标计算好了之后存储在数据仓库中，
4. 日志记录及警报发送。ETL 日志分为过程日志、错误日志、总体日志。记录日志的目的是随时可以知道 ETL 运行的可能错误。ETL 出错了，不仅要写 ETL 出错日志而且要向系统发送警报，并附上出错信息，方便排查错误。

3.2 指标设计

中国的银企仍然处于起步及快速增长阶段，银行分支机构的数量显著增加。银行分支机构数量在短时期内的迅速增加使得银企之间的竞争日益激烈；同时，银企还要和非银行机构进行竞争。另一方面，即便是对于历史悠久的欧洲银行业而言，无业绩贷款约占银企全部贷款的 55%-70%；在银行运作中，大多数银行有 10%-15% 的负扩张 (Zineldin, 1996)。这样，对银行来说，识别客户、维持客户已经越来越重要。银企已经被迫结束将重点放在吸引和补充新客户上的粗放性策略，转而将重点集中在发现、维持和发展有价值的客户上。

Zeithaml et al. (2001) 指出在银企系统，20% 的优质客户产生 82% 的银行零售利润；并且，发现银行以客户对银行贡献的利润率来分割客户。Blattberg and Deighton (1996) 认为，银企首先应该识别高价值客户并投资于具有最高价值的客户；Reichheld (1996) 认为，银企必须将他们的努力集中在能够连续向银企传递高价值的客户子集上，他建议银企应当通过考察下列问题来分离他们的关键客户：哪一客

户是最有利可图的和忠诚的、需要银行更少的服务、倾向于维持稳定和长期的关系的客户；哪一客户在银企所提供的产品和服务上贡献了最大价值；和银企的竞争对手相比，哪些客户对银企更有价值。基于银企已经建立的大型数据库，可以分析和应用于开发新的商务策略和机会。然而，不是同等地瞄准所有的客户或对所有客户都提供同样的激励，银企可以仅选择那些个人需要和购买行为符合一定利润标准的客户 (Dyché, 2002)。

直到目前为止，在有关的研究文献中，两个分数模型的建立一直基于注重实效的方法。因此，不存在对于每一个独特场合都适用的最好的分数模型。大多数研究将注意力放在建立更精确的信用或行为分数模型并使用各种统计技术增加分类模型的精确性。然而，由于银行数据库是多维的，由每月帐户记录和每天的交易记录构成，从而使用银行数据库分析客户行为是困难的 (Donato et al., 1999) 即使使用高度精确的分数模型，某些错分类情形仍频繁出现。为解决此问题，本项目试图借助于前沿的数据处理及挖掘技术，建立一个基于人工神经网络技术的客户识别模型进行客户关系管理方面的分析。

表 1: 银企客户数据库基本属性

	数据属性
数据库名	客户关系管理
数据来源	银企 ORACLE 数据库
样本容量	360,000,000,000 比特
指标类别	分类变量 30 个、数值变量 85 个
分析目的	预测客户资源的优劣

注意：CRM 数据的全部权限归于银企所有，本演示报告采用了模拟的客户关系数据。

图1显示了 CRM 项目数据的基本属性描述。从中可见，银企 ORACLE 数据库里包含了大量的字段以及海量的观察值。然而，这些数据并不可以直接使用来进行数据分析与挖掘。在进行实际的数据分析与挖掘之前，需要进行数据探查。图4、图5及图6为初步的数据探查图示。从图4可以看到，信用卡资金在 0-20 万之间的一般客户以及非持卡状态的客户的存款资金总量较多，并且前、后者之间的存款总量基本持平；信用卡处于欠款状态的客户资金总量较低，信用卡资金 21 万以上的充裕账户资金总量最少。从图5可见，对于高管人员、普通职工，信用卡资金在 0-20 万之间的或者不持卡的客户类型，存款资金总量最多；并且，其中的有房产者占了大部分比率。从图6可见，对于高管人员，信用卡资金在 0-20 万之间的或者不持卡的客户类型，存款资金平均值最高，其中的有房产者占了大部分比率；普通职工的情况类似，但是存款资金平均值要低于高管人员。这样，经过初步的数据探查后，可以获得如下能够实际应用于客户关系管理项目分析的部分指标：

属性 1: (分类)

信用卡资金状态：负值（欠款)、0-20 万之间（一般)、21 万以上（充裕）以及非持卡状态

属性 2: (数值)

贷款持续时间

属性 3: (分类)

贷款记录：当前是否持有贷款；信誉记录（本行的全部贷款按时偿还、全部贷款按时偿还、存在拖欠记录、持有其它银企信用卡的账户)

属性 4: (分类)

贷款目的：商业、住房、轿车、教育、装修、家电以及其它。

属性 5: (数值)

贷款数量

Attibute 6: (分类)

储蓄存款：不超过 10 万、11-50 万、51-100 万、101 万以上、未知或者无该类账户。

属性 7: (分类)

从业时间：失业、不超过 1 年、2-5 年、6-9 年、10 年以上

属性 8: (数值)

分期付款占可支配收入比率

属性 9: (分类)

性别及婚姻状态：男离异、女离异、男未婚、女未婚、男婚姻 (或丧偶)

属性 10: (分类)

是否存在其他债务人或者担保人：无、共同申请人、担保人

属性 11: (数值)

居住时间

属性 12: (分类)

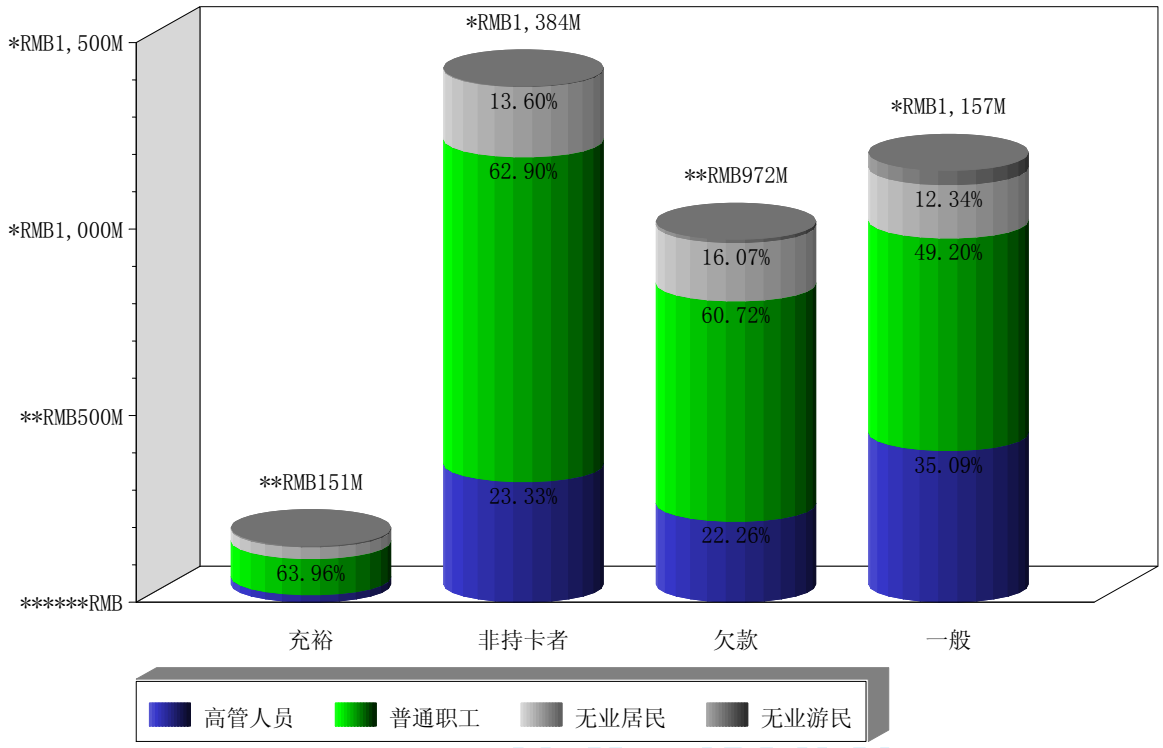
职业: 失业非居民、失业居民、技术工人或职员、官员或商人或高级职员。

.....

属性 20: (分类)

行代表实际分类和列为预测分类。如果类为劣的客户判断为类优的，则损失为 5；如果类为优良的客户判断为类劣的，则损失为 1。显然，前一种情况更糟糕。

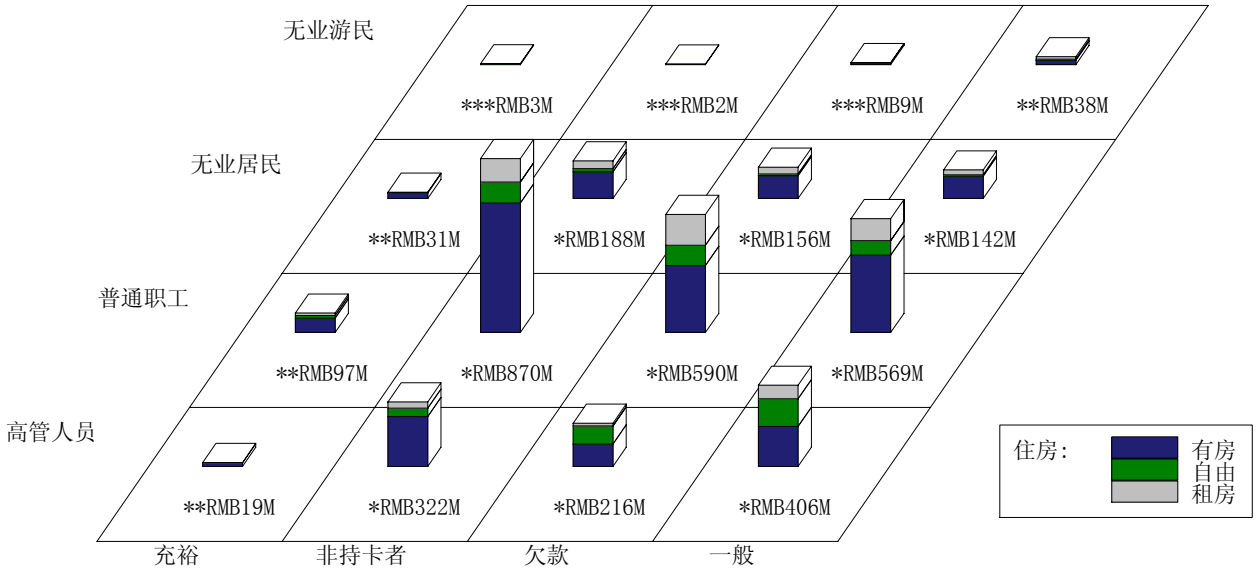
图 4: 信用卡资金状态 VS 持卡者职业及存款总量



注意：柱状图显示了不同资金状态的持卡者存款总量以及客户职业构成的百分比例。

CRM 项目数据的全部权限归银企所有，本演示报告采用了虚拟数据。

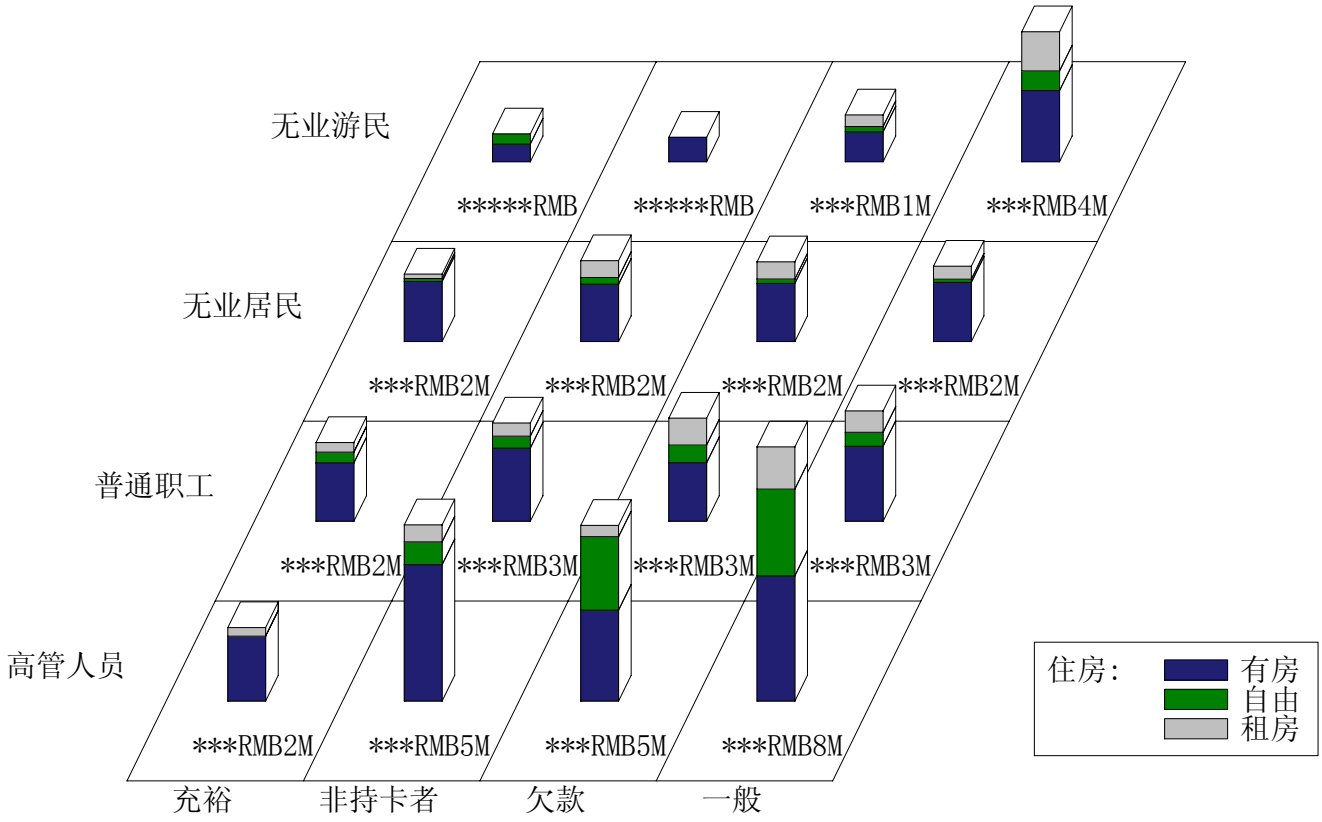
图 5: 信用卡资金状态、持卡者职业与房产以及存款总量



注意: 柱状图显示了信用卡资金状态、持卡者职业构成 VS 持卡者房产状态、账户存款总量。

CRM 项目数据的全部权限归银企所有，本演示报告采用了虚拟数据。

图 6: 信用卡资金状态、持卡者职业与房产以及存款均值



注意: 柱状图显示了信用卡资金状态、持卡者职业构成 VS 持卡者房产状态、账户存款均值。

CRM 项目数据的全部权限归银企所有，本演示报告采用了虚拟数据。

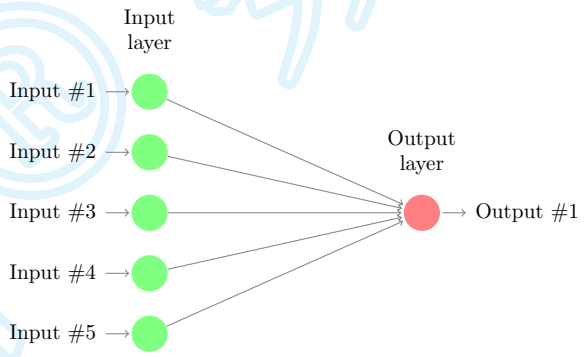
4 网络模型的估算与优选

基于高度简化的线性模型进行经济、金融系统的各类研究尝试，在多数情况下是困难而不切实际的。这就需要足够复杂的模型来拟合自变量与因变量之间的内在关系，避免拟合不足。神经网络模型便是进行复杂系统研究的最理想的工具之一。然而，该类模型复杂的计算过程不仅会让研究者望而却步，也极大地增加系统底层的开销。神经网络模型的复杂程度和隐藏单元的数目、各神经单元之间的连接数和参数值的大小有关。隐藏单元越多、连接数越多、参数绝对值越大，在模型越复杂。这样，无论简单的无隐藏层的 GLIM (Generalized LInear Model) 神经网络模型，还是复杂的多隐藏层的 MPL (Multi-Layer Perceptron)、ORBF (Ordinary Radial Basis Function) 或者 NRBF (Normalized Radial Basis Function) 模型，都应该在项目研究的考察范围之类。.

4.1 GLIM 模型

在经济、金融市场的研究中，Logistic 模型是最常用的一种工具。该模型的用途为：一、评估，比如上面所说的银企信用风险的评估；二、预测，在不同的指标变动条件下，预测产生信用风险的概率有多大；三、判别，类似于预测，即根据 logistic 模型，判断客户属于某种类型的概率有多大。Logistic 模型与线性回归之间存在着诸多相似之处，区别仅在于因变量设计形式不同，从而可以归于同一类模型，即 GLIM 模型。考察 GLIM 模型，可以看到：在因变量形式为单一线性的条件下，则为线性回归模型；在 Logistic 分布的前提下，则为 Logistic 回归；在 Poisson 分布时，就是 Poisson 回归，如此等等。在神经网络分析中，GLIM 模型被描述为一种具有前馈式结构的神经网络设计。这种神经网络模型没有隐藏层，并且输入与输出变量没有标准化，其结构类似于线性回归。

图 7: 无隐藏层的神经网络模型：GLIM



注意：无隐藏层，输入指标为 5 个、输出指标 1 个。

DMDB(数据挖掘的数据库) 是进行数据挖掘的前提。这就要求基于数据源建立 DMDB 数据库，基于指标角色编辑与计算元数据信息，并存储于元数据目录中。基于模型构建的需要，我们首先将全样本数据按照比例随机分为训练数据集和验证数据集两部分，并将数据信息存储于 DMDB 数据库中。此外，建立描述决策损失的矩阵如图2所示。我们将正确判断所产生的损失设定为 0，而将错误判断的损失设置为不同的正数值。比如，如果类为劣的客户判断为类优的，则损失为 5；如果类为优良的客户判断为类劣的，则损失为 1。显然，这是由于前一种情况产生的后果更严重。

表 2: 神经网络模型的决策矩阵设定

		决策的损失	
		To_1(预测为优)	To_2(预测为劣)
因变量观测值	=1(优良)	0	1
	=2(低劣)	5	0

根据样本数据的属性特征，将部分自变量指定为分类变量，其它部分指定为连续变量，再将因变量（客户资源优劣）设定为分类变量。在客户识别项目的 GLIM 模型里，组合函数被设计为线性，激活函数设计为 mlogistic 函数，误差函数为 mbernoulli 函数。这样，我们便可以基于 CRM 理论概念构建 GLIM 神经网络模型。图7显示了没有隐藏层的神经网络模型结构。

表 3: GLIM 模型的初始参数

Number	Parameter	GradObj	Estimate
1	var1A11var211	0.129714285	0
2	var1A12var211	0.094142857	0
3	var1A13var211	0.069285714	0
4	var3A30var211	0.050285714	0
5	var3A31var211	0.049714285	0
6	var3A32var211	0.048857142	0
7	var3A33var211	0.036142857	0
8	var4A40var211	0.011571428	0
9	var4A41var211	-0.024428571	0
10	var4A42var211	-0.011285714	0
11	var4A43var211	-0.029571429	0
12	var4A44var211	-0.009428571	0
13	var4A45var211	-0.008285714	0
14	var4A46var211	-0.002428571	0
15	var4A48var211	-0.011000000	0
16	var6A61var211	0.058714285	0
17	var6A62var211	0.023857142	0
18	var6A63var211	0.015428571	0
19	var6A64var211	0.012714285	0
20	var7A71var211	0.014285714	0
...	...	...	...
44	var11var211	-0.000714286	0
45	var13var211	-0.546857143	0
46	var16var211	-0.004000000	0
47	var18var211	-0.003428571	0
48	BIASvar211	2.88340E-16	0.8472

注意：目标函数值为 0.6108643021。目标函数定义为，样本容量除对应于各观测值的误差函数与惩罚函数之和。

表 4: GLIM 模型的最终参数

Number	Parameter	Estimate	GradObj
1	var1A11var211	-0.782258209	-2.574086E-6
2	var1A12var211	-0.392104431	-6.018491E-6
3	var1A13var211	0.059766314	-6.237578E-6
4	var3A30var211	-0.370328354	-4.672777E-6
5	var3A31var211	-1.020315468	-6.398489E-6
6	var3A32var211	0.035764048	3.028747E-7
7	var3A33var211	0.455660239	-5.470198E-6
8	var4A40var211	-0.730545887	5.233136E-6
9	var4A41var211	1.042916247	4.083580E-6
10	var4A42var211	0.431858501	7.330173E-7
11	var4A43var211	0.177260407	7.461973E-7
12	var4A44var211	-0.120145150	4.037721E-7
13	var4A45var211	-0.563166722	-1.661161E-7
14	var4A46var211	-0.635063513	3.778093E-7
15	var4A48var211	0.698566328	4.537621E-7
16	var6A61var211	-0.695530580	-5.168541E-7
17	var6A62var211	-0.065964539	-4.216013E-6
18	var6A63var211	-0.278661874	-4.023585E-6
19	var6A64var211	1.049958217	-3.208416E-6
20	var7A71var211	-0.620602271	-2.120786E-6
...	...	...	...
44	var11var211	-0.005139367	0.000029860
45	var13var211	0.020598777	0.000352999
46	var16var211	-0.428992157	0.000010994
47	var18var211	-0.206989449	9.938118E-6
48	BIASvar211	3.714696748	9.648465E-6

注意：目标函数值为 0.4437914972。目标函数定义为，样本容量除对应于各观测值的误差函数与惩罚函数之和。

4.2 MLP 模型

MLP 模型是最常见的一种神经网络模型，其隐藏层的组成通常为线性组合函数和 S 型激活函数，输出层的则使用线性组合函数和与因变量相适应的激活函数。不失一般性，我们把客户识别模型设计为具有线性组合函数的隐藏层、tanh (hyperbolic tangent) 激活函数的输入层、恒等函数或者无激活函数的输出层的 MLP 神经网络模型结构。在这样的设定下，某个隐藏层的神经元便能够以公式表示为

$$H_i = \tanh\left(-b_j^2 \sum w_{ij} x_i\right)$$

这里， x_i 是第 i 个神经元的输入值， w_{ij} 是第 i 个输入神经元至第 j 个隐藏神经元的权重， b_j 为第 j 个隐藏神经元的偏离。

MLP 模型被称为通用拟合机制 (Universal Approximator)。在给予足够数据、隐藏神经元和训练时间的条件下，包含单隐藏层的 MLP 模型能够以任意精度拟合自变量与因变量之间的几乎任意形式的函数。尽管如此，使用更多的隐藏层则可以减少隐藏神经元 (从而减少参数) 的数量，提高模型的普遍适用性。进一步，为了避免模型过于复杂、减少噪音学习和过度拟合，我们可以采用早停止法 (Early Stopping)、权重衰减 (Weight Decay) 以及网络修剪等方法来获得相对简化的模型。

表5列出了基于不同类型的神经网络模型对于验证集进行预测所获得的统计量，其中的 MLP 模型都设置为单隐藏层的结构。通过比较各类模型对于验证集预测的总损失，我们发现使用早停止法建立的具有两个隐藏神经元的 MLP 模型的决策损失为 148，低于其它模型。因此，该模型的预测效果最好。

表 5: 各神经网络模型对于验证集的预测效果

总损失	隐藏神经元	衰减系数	VMAX	VMSE	VRASE	VRMSE	VMISC	VALOSS	架构	增强推广方法
162	0	0.0000	0.973	0.159	0.399	0.399	0.217	0.540	GLIM	
153	1	0.0000	0.963	0.171	0.413	0.413	0.253	0.510	MLP	Early Stopping
201	1	0.1000	0.914	0.181	0.426	0.426	0.257	0.670	MLP	Weight Decay
218	1	0.0100	0.945	0.194	0.441	0.441	0.280	0.727	MLP	Weight Decay
187	1	0.0010	0.947	0.185	0.431	0.431	0.277	0.623	MLP	Weight Decay
191	1	0.0001	0.954	0.191	0.437	0.437	0.287	0.637	MLP	Weight Decay
149	2	0.0000	0.969	0.170	0.412	0.412	0.253	0.497	MLP	Early Stopping
202	2	0.1000	0.998	0.222	0.471	0.471	0.300	0.673	MLP	Weight Decay
248	2	0.0100	1.000	0.234	0.484	0.484	0.327	0.827	MLP	Weight Decay
251	2	0.0010	1.000	0.226	0.475	0.475	0.303	0.837	MLP	Weight Decay
223	2	0.0001	0.999	0.215	0.464	0.464	0.293	0.743	MLP	Weight Decay
155	3	0.0000	0.955	0.162	0.402	0.402	0.217	0.517	MLP	Early Stopping
180	3	0.1000	1.000	0.231	0.480	0.480	0.297	0.600	MLP	Weight Decay
171	3	0.0100	1.000	0.225	0.475	0.475	0.280	0.570	MLP	Weight Decay
178	3	0.0010	1.000	0.237	0.487	0.487	0.293	0.593	MLP	Weight Decay
177	3	0.0001	1.000	0.226	0.475	0.475	0.277	0.590	MLP	Weight Decay

注意：表中的 MLP 神经网络模型都是单隐藏层的设计结构。

表 6: MLP 神经网络模型的初始参数

序号	参数名	梯度	估计值
1	var1A11H11	0.008024785	0.768970072
2	var1A12H11	0.001288708	0.120935428
3	var1A13H11	0.002612726	-0.253851946
4	var3A30H11	-0.005650012	0.369071848
5	var3A31H11	-0.004019668	1.002281935
6	var3A32H11	-0.008228825	0.062756312
7	var3A33H11	-0.003081673	-0.640710582
8	var4A40H11	0.005727757	0.598976518
9	var4A41H11	0.002761358	-0.587405593
10	var4A42H11	0.005212648	-0.440357921
11	var4A43H11	0.004550693	-0.138514656
12	var4A44H11	0.003575137	0.167635888
13	var4A45H11	0.002547789	0.110920129
14	var4A46H11	-0.001528953	0.248015323
15	var4A48H11	0.003928617	-0.125910047
16	var6A61H11	-0.005715263	0.371651748
17	var6A62H11	-0.000082965	0.303267017
18	var6A63H11	-0.004946862	-0.004590245
19	var6A64H11	-0.000711237	-0.447502919
20	var7A71H11	0.001427055	0.763438594
...	...	...	...
95	BIASH11	0.000082891	-0.157832515
96	BIASH12	0.000764304	-0.134793123
97	H11var211	-0.000022471	-1.395852835
98	H12var211	-0.016538932	1.199677475
99	BIASvar211	0.015116363	1.400342314

注意：本表报告了具有两个隐藏神经元的单隐藏层 MLP 模型的参数估计。

表 7: MLP 神经网络模型的最终参数

Number	Parameter	Estimate	GradObj
1	var1A11H11	0.768970072	0.008024785
2	var1A12H11	0.120935428	0.001288708
3	var1A13H11	-0.253851946	0.002612726
4	var3A30H11	0.369071848	-0.005650020
5	var3A31H11	1.002281935	-0.004019668
6	var3A32H11	0.062756312	-0.008228825
7	var3A33H11	-0.640710582	-0.003081673
8	var4A40H11	0.598976518	0.005727757
9	var4A41H11	-0.587405593	0.002761358
10	var4A42H11	-0.440357912	0.005212648
11	var4A43H11	-0.138514651	0.004550693
12	var4A44H11	0.167635888	0.003575137
13	var4A45H11	0.110920129	0.002547789
14	var4A46H11	0.248015323	-0.001528953
15	var4A48H11	-0.125910047	0.003928670
16	var6A61H11	0.371651748	-0.005715263
17	var6A62H11	0.303267017	-0.000082965
18	var6A63H11	-0.004590245	-0.004946862
19	var6A64H11	-0.44750299	-0.000711237
20	var7A71H11	0.763438594	0.001427055
...	...	...	...
95	BIASH11	-0.157832515	0.000082891
96	BIASH12	-0.134793103	0.000764304
97	H11var211	-1.395852835	-0.000022471
98	H12var211	1.199677475	-0.016538932
99	BIASvar211	1.400342315	0.015116363

注意：本表报告了具有两个隐藏神经元的单隐藏层 MLP 模型的参数估计。

4.3 效率评估

数据挖掘的全过程包括了抽样、探索、修改、建模和评估等环节，这些都是为了发现关联因素以及未知模式。抽样过程为确认分析数据集，将数据分割为训练集、验证集和测试集。探索过程则为以描述性统计和图形的方式，熟悉数据、寻找不寻常的事件或观测。修改过程为准备分析数据（创建新指标、修改旧指标、处理奇异值、替换缺失值等)。建模则为，使用回归、决策树、神经网络或者其它模型理论，模型化自变量与因变量之间关系。最后，评估过程。这是为了比较不同类型预测模型的使用效率。

表 8: MLP 神经网络模型的预测效率：基于训练集的判断

观测序号	因变量值	决策	判为 1 概率	判为 2 概率	误差函数 1	误差函数 2	误判损失
1	1	To_1	0.971	0.029	0.060	0	0
2	1	To_1	0.981	0.019	0.039	0	0
3	1	To_2	0.801	0.199	0.443	0	1
4	1	To_1	0.973	0.027	0.054	0	0
5	1	To_2	0.72	0.280	0.657	0	1
6	1	To_2	0.629	0.371	0.928	0	1
7	1	To_2	0.323	0.677	2.262	0	1
8	1	To_1	0.968	0.032	0.066	0	0
9	1	To_1	0.846	0.154	0.335	0	0
10	1	To_1	0.951	0.049	0.101	0	0
11	1	To_1	0.968	0.032	0.065	0	0
12	1	To_1	0.954	0.046	0.095	0	0
13	1	To_1	0.909	0.091	0.192	0	0
14	1	To_1	0.934	0.066	0.136	0	0
15	1	To_2	0.726	0.274	0.641	0	1
16	1	To_2	0.824	0.176	0.388	0	1
17	1	To_2	0.506	0.494	1.360	0	1
18	1	To_2	0.468	0.532	1.521	0	1
19	1	To_1	0.874	0.126	0.270	0	0
20	1	To_2	0.800	0.200	0.446	0	1
21	1	To_2	0.647	0.353	0.870	0	1
22	1	To_2	0.82	0.180	0.396	0	1
23	1	To_1	0.907	0.093	0.196	0	0
24	1	To_1	0.939	0.061	0.126	0	0
25	1	To_2	0.783	0.217	0.489	0	1
		...		...		...	
696	2	To_2	0.751	0.249	0	2.777	0
697	2	To_2	0.371	0.629	0	0.927	0
698	2	To_1	0.85	0.15	0	3.799	5
699	2	To_2	0.676	0.324	0	2.252	0
700	2	To_2	0.32	0.68	0	0.77	0

注意：具有两个隐藏神经元的单隐藏层 MLP 模型的预测结果。To_1 表示决策为 1（优质)，To_2 决策为 2（劣质)；误差函数 1 对应优质客户误判的情形；误差函数 2 则为劣质客户误判的情形。

表8和表9显示了 MLP 神经网络模型的预测结果，分别对应着训练集、验证集。从这些表格可以看到，无论是训练集、验证集，将优质客户判断为类劣的错误频率较高，但将劣质客户资源判断为类优的错误则很少出现。显然，这两类错误的成本之间存在着某种置换 (tradeoff)。

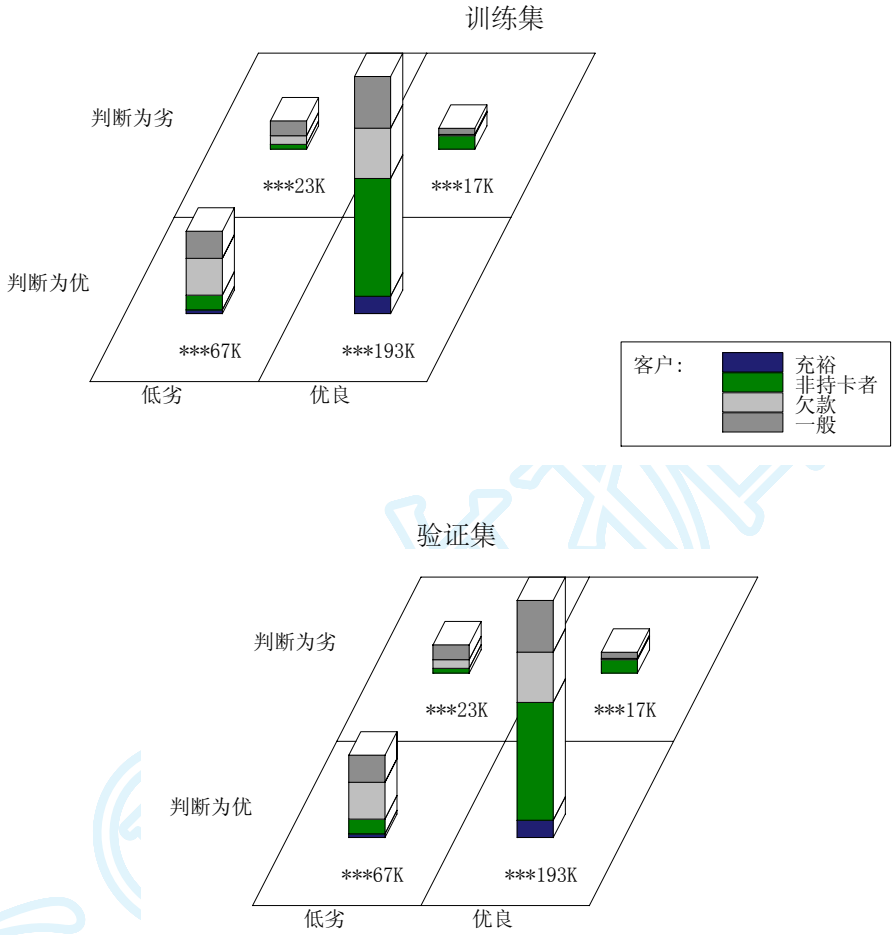
表 9: MLP 神经网络模型的预测效率：基于验证集的判断

观测序号	因变量值	决策	判为 1 概率	判为 2 概率	误差函数 1	误差函数 2	误判损失
1	1	To_2	0.830	0.170	0.374	0	1
2	1	To_1	0.981	0.019	0.038	0	0
3	1	To_2	0.413	0.587	1.766	0	1
4	1	To_1	0.961	0.039	0.079	0	0
5	1	To_1	0.932	0.068	0.141	0	0
6	1	To_1	0.914	0.086	0.181	0	0
7	1	To_1	0.974	0.026	0.052	0	0
8	1	To_2	0.437	0.563	1.657	0	1
9	1	To_2	0.826	0.174	0.382	0	1
10	1	To_1	0.883	0.117	0.249	0	0
11	1	To_1	0.840	0.16	0.349	0	0
12	1	To_1	0.913	0.087	0.182	0	0
13	1	To_2	0.802	0.198	0.441	0	1
14	1	To_1	0.944	0.056	0.114	0	0
15	1	To_2	0.802	0.198	0.442	0	1
16	1	To_1	0.838	0.162	0.355	0	0
17	1	To_2	0.788	0.212	0.478	0	1
18	1	To_2	0.590	0.410	1.054	0	1
19	1	To_2	0.464	0.536	1.537	0	1
20	1	To_2	0.827	0.173	0.379	0	1
21	1	To_1	0.937	0.063	0.130	0	0
22	1	To_1	0.979	0.021	0.042	0	0
23	1	To_1	0.876	0.124	0.265	0	0
24	1	To_1	0.921	0.079	0.165	0	0
25	1	To_1	0.970	0.030	0.062	0	0
		...		...		...	
296	2	To_2	0.488	0.512	0	1.34	0
297	2	To_2	0.772	0.228	0	2.961	0
298	2	To_1	0.95	0.05	0	5.999	5
299	2	To_2	0.239	0.761	0	0.546	0
300	2	To_2	0.744	0.256	0	2.728	0

注意：具有两个隐藏神经元的单隐藏层 MLP 模型的预测结果。To_1 表示决策为 1（优质），To_2 决策为 2（劣质）；误差函数 1 对应优质客户误判的情形；误差函数 2 则为劣质客户误判的情形。

图8显示了 GLIM 模型 (Logistic 模型) 的预测效率。从图8可以看到，虽然 GLIM 模型将优质客户判断为类劣的错误频率极低，但将劣质客户资源判断为类优的误判频率极高 (训练集、验证集的误判比率大约为 7 成或 8 成)。这表明，基于这种客户识别模式，为了增加优质客户的数量，需要以接纳更多劣质客户为代价。因此，可以认为广义线性模型的决策代价是极高的。

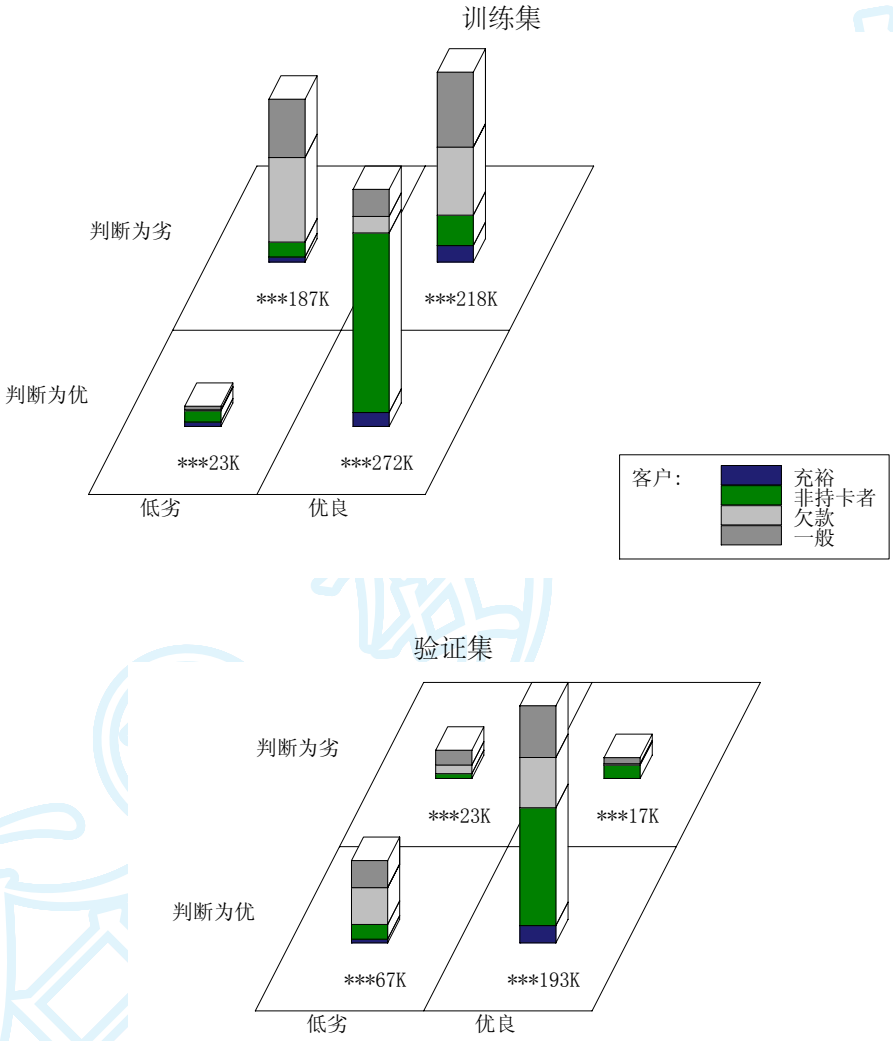
图 8: GLIM 模型 (Logistic 模型) 预测效率



注意：在训练集中，优质客户资源数量为 490K，被误判为劣质的比率为 9.39%；劣质客户资源 210K，误判为优质的比率为 76.67%。在验证集中，优质客户资源数量为 210K，被误判为劣质的比率为 8.10%；劣质客户资源 90K，误判为优质的比率为 74.44%。

图9显示了具有两个隐藏神经元的单隐藏层 MLP 模型的预测效率。从图9可以看到，MLP 模型将优质客户判断为类劣的频率较高 (训练集、验证集的误判比率约为 4 成)，但将劣质客户资源判断为类优的误判频率极低 (训练集、验证集的误判比率约为 1 成)。这表明，基于这种客户识别模式，可以精确判断优质客户，拒绝劣质客户对于银企资源的耗费。虽然这是以牺牲部分优质客户资源为代价的，但根据前面部分的决策矩阵成本设定，却是实现了最优化的成本付出。因此，这种单隐藏层的 MLP 神经网络模型的决策方式是比较理想的。

图 9: MLP 神经网络模型预测效率



注意：具有两个隐藏神经元的单隐藏层 MLP 模型的预测效率。在训练集中，优质客户资源数量为 490K，被误判为劣质的比率为 44.49%；劣质客户资源 210K，误判为优质的比率为 10.95%。在验证集中，优质客户资源数量为 210K，被误判为劣质的比率为 42.38%；劣质客户资源 90K，误判为优质的比率为 13.33%。

5 总结

客户资源的识别与管理是银企一项长期而复杂的任务。如何识别和预防欺诈性客户、拖欠贷款客户却一直是银企风险管理的一个难题。银企需要根据掌握的客户信息，使用数据挖掘技术识别优质客户资源。识别优质客户的关键则在于定量化客户对组织的价值，并在此基础上精确判断场客户群、寻找潜在客户和目标市场，采取积极措施吸引、维持客户。

5.1 项目成绩

基于银企业务行为的视角，客户关系管理不仅成为了银行日常工作的重要内容，也是防范坏账风险的关键环节。客户风险可能在整个关系过程中发生显著变化，有效的 CRM 管理能够使得银企以竞争性价格为客户群体提供优质的产品和服务，同时减少经营成本和风险损失。通过对于客户数据进行深入挖掘，一方面能够帮助银企科学评价优质客户资源获取的关键因素；进一步地，通过了解客户价值驱动的根本因素，银企可以对客户进行排序和管理客户并优化收入和支付，识别和拒绝欺诈性客户。

项目组通过对银企客户识别和客户保持问题的定量化建模研究与分析，深入探讨了客户关系管理的策略制定方法；同时，对 CRM 管理中的主要问题进行了分析，通过实例检验了部分模型的可行性，获得了符合管理实践的相应的客户关系管理策略。

1. 构建了银企 CRM 管理的指标体系。根据银企 ORACLE 数据库的字段类型及样本容量，采用企业级 ETL 技术对于数据源进行了有效处理。通过探查数据，建立了一套适于银企的指标体系。在大型数据库中，使用数据挖掘技术，分析银企的客户数据资料。从而，为银企完善 CRM 管理体系，提供了必要的前提和基础。
2. 设计了客户识别的神经网络模型系统。提出了基于单隐藏层 MLP 神经网络模型的客户识别模型。该模型可以有效地识别银企的有价值客户，防范欺诈客户。为银企判断、维持和发展客户，提供了科学、客观、有效的判断标准。进一步，这也为银行的融资、投资等核心业务的决策行为提供了可靠的技术支持。

在研究方法上，我们将 CRM 理论同先进的数据挖掘方法、系统决策理论、优化建模方法和经济分析方法结合起来，实现了管理、经济理论与计量、统计理论的融合，定性分析与定量建模方法的统一。

5.2 进一步研究

客户识别和维持问题是一个较新的研究领域。由于客户需求的快速多变和客户选择的空间增大，识别优质客户资源和保持有价值的、忠诚的客户，成为银企增强竞争力、降低风险、增加收入的关键。这一领域已经成为 CRM 研究的热点。实践发展领先于理论研究，但实践中的管理措施需要从理论上抽象，分析其合理的因素以及在实践中可行的管理策略。在 CRM 决策与优化研究方面，还存在诸多值得进一步研究的问题：基于动态个体客户利润率的客户识别模型研究；使用信用分数模型的错分类模式识别拖欠贷款客户的研究；与中、小企业的银行转移有关的因素和模型研究；使用比例风险模型对银行服务进行客户流失分析和模型研究；基于客户购买量的客户识别模型研究。这些问题都能够基于神经网络模型得出一般性的规律，从而辅助 CRM 管理的决策与优化。项目组将继续对上述领域作进一步探索，期望获得兼具理论和应用价值的研究成果。

参考文献

Blattberg, R.C. and J. Deighton, 1996. Manage Marketing by the customer equity test. *Harvard Business Review.* 74:136–44.

Chen, Mu-Chen and Shih-Hsien Huang, 2003. Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications.* 24:433 – 441

Donato, June M., Jack C. Schryver, Gregory C. Hinkel, Richard L. Schmoyer Jr., Michael R. Leuze and Nancy W. Grandy, 1999. Mining multi-dimensional data for decision support. *Future Generation Computer Systems.* 15:433 – 441

Dyché, Jill, 2002. *The CRM handbook: a business guide to customer relationship management.* Addison-Wesley

Haefke, Christian and Christian Helmenstein, 1996. Forecasting Austrian IPOs: An application of linear and neural network error-correction models. *Journal of Forecasting.* 15:237–251

Hagan, Martin T., Howard B. Demuth and Mark H. Beale, 2002. *Neural Network Design*

Kim, Yoon Seong and So Young Sohn, 2004. Managing loan customers using misclassification patterns of credit scoring model. *Expert Systems with Applications.* 26:567 – 573

Kodogiannis, V. and A. Lolis, 2002. Forecasting Financial Time Series using Neural Network and Fuzzy System-based Techniques. *Neural Computing & Applications.* 11:90–102

Kuan, Chung-Ming and Tung Liu, 1995. Forecasting Exchange Rates Using Feedforward and Recurrent Neural Networks. *Journal of Applied Econometrics.* 10:347–64

Kuan, Chung-Ming and Halbert White, 1994. Artificial neural networks: an econometric perspective. *Econometric Reviews.* 13:1–91

Lisi, F. and R.A. Schiavo, 1999. A comparison between neural networks and chaotic models for exchange rate prediction. *Computational Statistics and Data Analysis.* 30:87–102

Malhotra, Rashmi and D.K. Malhotra, 2003. Evaluating consumer loans using neural networks. *Omega.* 31:83 – 96

Reichheld, Frederick F., 1996. Learning from customer defections. *Harvard Business Review.* 74:56–69

Setiono, Rudy, James Y.L Thong and Chee-Sing Yap, 1998. Symbolic rule extraction from neural networks: An application to identifying organizations adopting IT. *Information & Management.* 34:91 – 101

Thomas, Lyn C., 2000. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting.* 16:149 – 172

White, Halbert, 1988. Economic prediction using neural networks: the case of IBM daily stock returns. In: *IEEE International Conference on Date of Conference: 24-27 July 1988.* 2:451 – 458. Dept. of Econ., California Univ., San Diego, CA, USA

Zeithaml, Valarie A., Roland T. Rust and Katherine N. Lemon, 2001. The Customer Pyramid: creating and serving profitable customers. *California Management Review.* 43:118–142

Zineldin, Mosad, 1996. Bank strategic positioning and some determinants of bank selection. *International Journal of Bank Marketing.* 14:12–22

Thursday, March 15, 2012 15:34:42

人大经济论坛评谷®数据处理和分析系列项目报告

人大经济论坛评谷® 数据处理与分析中心
电话: 86+010-684 542 76
地址: 北京市海淀区厂洼街 3 号
网址: http://data.pinggu.org