

基于多进化神经网络的信用评估模型研究

喻敏¹ 吴江^{1,2}

(西南财经大学经济信息工程学院 成都 610074)¹ (四川大学计算机学院 成都 610065)²

摘要 客户信用评估对于银行的经营管理有着重要的意义,为此提出了一种基于多进化神经网络的信用评估模型(MNN-CREDIT)。该模型基于客户信贷数据,利用基于聚类的小生境遗传算法并行地训练出多个精度高、差异性大的三层前馈神经网络,然后将待识别的客户数据分别输入,最后根据动态投票法集成最终信用预测结果。利用德国信用数据库真实数据集进行了实证分析,结果表明,基于多进化神经网络的信用评估模型具有较高的预测精度。

关键词 信用评估,小生境遗传算法,神经网络

中图分类号 F833.2 **文献标识码** A

Credit Evaluation Model Based on Multiple Evolutionary Neural Networks

YU Min¹ WU Jiang^{1,2}

(School of Economic Information Engineering, Southwest University of Finance and Economics, Chengdu 610074, China)¹

(School of Computer Science, Sichuan University, Chengdu 610065, China)²

Abstract Credit evaluation plays an important role in the banking management. A novel credit evaluation model based on multiple evolutionary neural networks, named MNN-CREDIT, was presented. The MNN-CREDIT model establishes classifiers by a group of three-layer feed-forward neural networks with high accuracy and good diversity. The neural networks are trained by niche genetic algorithm based on clustering. The credit evaluation result of the identifying client can first be evaluated by each neural network, and the final credit classification result is obtained according to the dynamic voting rule. Empirical analysis on Germany credit database was given. The results show that MNN-CREDIT model has higher prediction precision.

Keywords Credit evaluation, Niche genetic algorithm, Neural network

1 引言

信用评估是商业银行控制风险的关键技术。随着中国经济的快速发展,消费信贷热不断升温,各商业银行均把个人贷款业务作为未来发展战略的重要组成部分。目前,尚缺乏一套有效的个人信用评估方法,阻碍了个人消费信贷业务的进一步开展。因此建立一个快速、高效的信用评估模型,对金融市场的发展具有重要意义。

信用评估实质上是模式识别的分类问题——将贷款者划分为能够按期还本付息的“好”客户和违约的“坏”客户两类。训练时,根据历史上每个类别(按期还本付息和违约)的若干样本,从中找出违约和不违约者的特征,从而得到分类规则,并建立数学模型,进而预测未来贷款人的违约风险,为消费信贷决策提供科学依据^[1]。

针对信用评估对象的特点,本文提出了基于多进化神经网络的信用评估模型(MNN-CREDIT)。利用该模型对德国信用数据库的真实数据进行实证研究,取得了较高的预测精度。结果表明 MNN-CREDIT 模型较朴素贝叶斯模型,其预测精度提高了 3.04%;较支持向量机,其预测精度提高了 1.64%;

较 k -近邻判别分析方法($k=17$),其预测精度提高了 6.87%。

2 文献综述

传统的信用评估方法有统计评估和非统计评估两大类。统计评估方法主要包括判别分析、线性回归、非线性回归、Logit 模型以及非参数统计中的 k -近邻判别分析方法(KNN)等。非统计评估方法包括线性规划、整数规划、神经网络、专家系统等^[2]。

判别分析模型形式简单直观,具有很好的解释性,但必须满足现实中难以满足的等协方差和正态等条件。二次判别分析模型可解决等协方差阵条件,但却不满足正态假定,并且当数据样本维度高、数量少时性能显著下降。这两种方法必须假设样本满足某种统计分布,因此显得过于牵强。Logit 模型无需假定任何概率分布,也不要求等协方差性,但当样本点存在完全分离时,模型参数的最大似然估计可能不存在。另外该方法对中间区域的判别敏感性较强,导致判别结果不稳定^[2]。KNN 方法是一种常用的非参数模式识别方法,它不要求数据正态分布,主要特点在于它的非参数的特点使得在特征变量空间上对于不规则变量的建模成为可能。然而当数据

到稿日期:2010-10-15 返修日期:2011-01-13 本文受西南财经大学科研基金项目(QN0806),西南财经大学“211工程”三期青年教师成长项目(211QN09071)资助。

喻敏(1961—),女,讲师,主要研究方向为数据挖掘、商务智能。

维数比较高时,存在所谓的“维数灾”问题,且即使样本量很大,其散落在高维空间中仍显得非常稀疏,绝大多数点附近根本没有样本点,这使得 KNN 的思想很难体现^[3]。

人工神经网络是一种对数据分布无任何要求的非线性技术,它能有效解决非正态分布、非线性的信用评估问题,但是单个神经网络易出现过学习现象,推广能力较差^[2]。

可见,本质上信用评估是一个复杂的非线性分类问题,因此使用线性模型进行分类和预测的效果并不理想;而非线性统计和智能方法也都各有利弊。目前还没有一个公认的最优信用评估模型,因此探索新的信用评估模型和方法具有重要的理论价值和实际意义。

3 MNN-CREDIT 模型信用评估流程

构建的 MNN-CREDIT 信用评估模型具有以下 4 个模块。图 1 描述了信用评估流程。

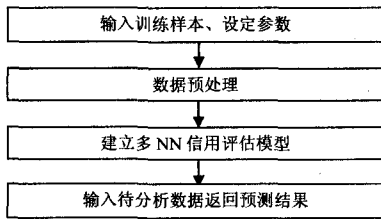


图 1 MNN-CREDIT 信用评估流程

1. 用户交互模块:负责设定相关参数,输入数据。
2. 数据预处理模块:负责数据的清理和表示工作。
3. 评估模型训练模块:利用多神经网络集成根据输入数据挖掘信用评估模型。
4. 信用预测模块:根据得到的 MNN-CREDIT 信用评估模型预测待分析记录。

4 MNN-CREDIT 信用评估模型

理论和实证表明,为了提高最终的预测精度,在保证所训练神经网络的泛化能力的同时,还应该提高它们之间的差异^[4]。因此,MNN-CREDIT 采用适合多模函数优化问题的小生境遗传算法,同时进化出多个精度高且差异性大的神经网络。MNN-CREDIT 信用评估模型包括两个步骤:(1) 采用结合聚类的小生境遗传算法对多个三层前馈神经网络进行训练;(2) 综合各个神经网络的信用评估结果对待识别客户进行信用评估。MNN-CREDIT 模型的基本框架如图 2 所示。

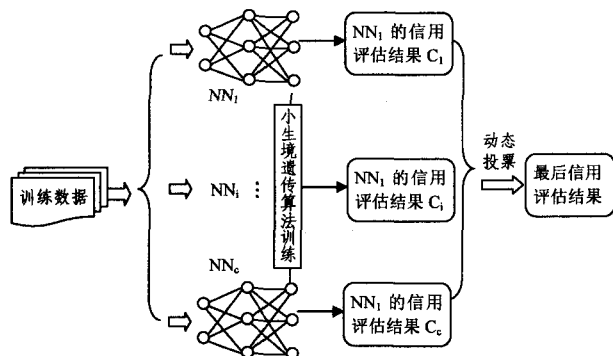


图 2 MNN-CREDIT 信用评估模型

4.1 数据预处理

现实生活中收集的信用数据一般存在空缺数据、不一致

数据,并且表达形式多样化,不利于进一步的数据挖掘。为提高挖掘结果的精度和有效性,必须对数据进行预处理。在 MNN-CREDIT 中,数据预处理分两步进行:

1) 数据清理。这一步用于处理原始数据集中存在的空缺数据或奇异值数据,要点包括:

- a) 对样本数据中有缺失属性值的样本进行删除。
- b) 对每个属性的奇异值用领域专家认定的相应的缺省值来代替。

2) 数据表示。原始数据的形式可能不利于数据分析,因此需要对数据进行离散化和概念分层,要点包括:

- a) 离散型数据:用 1 到 N 之间的离散型数据表示各属性值。离散型变量只取集合内的固定值。
- b) 连续型数据:

采用比例变换的方法:某些属性的值与其他属性的值相比,数值相差太大,会造成分类器性能下降,需要通过线性比例变换映射到某一区间。例如,贷款金额一般在区间 $[1000, 9999]$,可以通过比例变换映射到区间 $[10, 99]$ 。

通过划分分段:根据属性值的范围划分为 N 个不相交的区间,分别用编号代替属于每个区间的属性的值。例如,把工作时间表示成如下形式:1. 失业;2. 小于 1 年;3. 1~4 年;4. 4~7 年;5. 7 年以上。

4.2 神经网络模型的训练

4.2.1 染色体编码

针对神经网络的结构,染色体编码将网络连接权值和节点的偏值进行混合实数编码。码串长度为 $l = (n_i + n_o)n_h + n_h + n_o$,其中 n_i 为输入层节点个数, n_o 为输出层节点个数, n_h 为隐层节点个数。编码的具体方式为:按照输入层和隐层节点间连接权→隐层和输出节点间连接权→隐层节点偏值→输出节点偏值的顺序把各权值和偏值连成一串。图 3 为一个 3 层神经网络的染色体编码示例。该网络有两个输入神经元、3 个隐层神经元和一个输出神经元。编码为:4. 25, 1. 02, -2. 93, 0. 85, 2. 74, -3. 81, 3. 45, 0. 9, -2. 4, 0. 8, 1. 5, -0. 75, 0. 5。

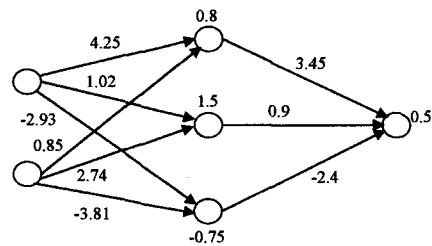


图 3 神经网络的染色体编码

4.2.2 适应度函数的设计

适应度函数用于衡量染色体的优劣,即神经网络对样本数据的逼近精度和网络的泛化能力。在本模型中,适应度公式为:

$$f_i = \frac{1}{\left(\sum_{j=1}^n (C_{i,j} - T_j) / n \right) + 1} \quad (1)$$

其中

$$C_{i,j} - T_j = \begin{cases} 0, & C_{i,j} = T_j \\ 1, & \text{otherwise} \end{cases} \quad (2)$$

式中, $C_{i,j}$ 表示第 i 个个体代入第 j 个样本后得到的分类结果(类标); T_j 表示第 j 个样本中实际的类标; n 则为样本数量。

显然,当 $C_{i,j} = T_j$ ($j=1,2,\dots,n$) 时,适应度函数的取值将最大,即 f_i 的最大值为 1。

4.2.3 初始化种群

首先依据经验确定种群中连接权值和偏值的取值范围。初始化种群时,连接权值和偏值使用取值范围内的随机数,以保证初始种群中连接权值和偏值的多样性和均匀分布性。

4.2.4 交叉算子

采用两点交叉算子,以概率 P_c 交叉两个个体,父辈染色体 $f_1=[w_1,\dots,w_n,v_1,\dots,v_m]$, $f_2=[w_1',\dots,w_n',v_1',\dots,v_m']$,其中 w_i 为连接权, v_i 为偏值。随机生成与其等长的两个二进制模板 t_1 和 t_2 。如果模板 t 中的一位为 0,表示由该模板生成的染色体的该位基因来自父辈染色体 f_1 ;如果为 1,则表示由该模板生成的染色体的该位基因来自父辈染色体 f_2 。交叉算子示例如图 4 所示。

$$\begin{array}{ccccccc} f_1=[w_1,\dots,w_n,v_1,\dots,v_m] & f_2=[w_1',\dots,w_n',v_1',\dots,v_m'] \\ | & \dots & | & & | & \dots & | \\ t_1=[1 & \dots & 1 & 1 & \dots & 0] & t_2=[0 & \dots & 1 & 0 & \dots & 0] \\ \downarrow & & & & \downarrow & & \\ f_1'=[w_1',\dots,w_n',v_1',\dots,v_m] & f_2'=[w_1,\dots,w_n,v_1,\dots,v_m] \end{array}$$

图 4 交叉算子示例

4.2.5 变异算子

变异算子为:对选中变异的基因 ξ 加上一个随机的偏置值 r 。连接权值以及偏值变异时,依照经验将基因位的取值限定在 $[a,b]$ 范围内,因此取 r 为 $[a,b]$ 之间的一个随机值,按照 $\xi'=(\xi+r)/2$ 对染色体中选中的基因 ξ 做变异。此变异保证了变异后 ξ' 的取值仍然在限定区间 $[a,b]$ 中。

4.2.6 选择算子

采用轮盘赌选择和精英保留策略,个体的适应度越高,被复制到下一代的份数就越多。各个体在下一代群体中的期望生存数目为:

$$N_i = N \cdot f_i / \sum_{j=1}^N f_j \quad (i=1,2,\dots,N) \quad (3)$$

式中, N 为种群中个体总数。Radolph 证明了一般的遗传算法不一定收敛,只有每代保存了最优个体时才收敛^[5]。因此算法将当前适应度最高的基因另外保存一份,以免其在交叉或变异中被破坏掉。若发现比当前保存的基因适应度更高的基因,则用其替换当前的保存。

4.2.7 子神经网络间的距离

在 MNN-CREDIT 中,种群中的每个个体就是一个神经网络,采用式(4)去衡量个体 NN_i 和 NN_j 之间的距离^[6]。

$$D(NN_i, NN_j) = \frac{1}{N} \sum_{k=1}^N Diff(C_i(x_k), C_j(x_k)) \quad (4)$$

式中, N 是训练数据集中样本的个数, $C_i(x_k)$ 是第 i 个神经网络对第 k 个样本所得的分类结果,即类标。其中,

$$Diff(C_i(x_k), C_j(x_k)) = \begin{cases} 0, & C_i(x_k) = C_j(x_k) \\ 1, & \text{otherwise} \end{cases} \quad (5)$$

个体间的距离介于 0 到 1 之间。当对所有训练样本,两个神经网络的分类结果相同时,则它们之间的距离等于 0;反之,当两个神经网络的分类结果全部不同时,则它们之间的距离等于 1。

4.2.8 基于聚类的小生境技术

在计算完每两个个体之间的距离之后,对种群中的个体进行聚类并对每类中除适应度最高的个体外的其他个体执行罚函数。其目的是进行多模函数优化,得到多个适应度高且

差异性大的个体(神经网络)。MNN-CREDIT 使用 K-Means 算法对种群中的个体进行聚类,把种群划分成 c 个类。在每次的遗传迭代中,算法组合当前种群和上一代中最好的 c 个个体并且使用 K-Means 算法把整个种群聚类成 c 个类。在每个类中,除了适应度最高的个体外,其他个体对其适应度施加一个罚函数,个体 NN_i 施加罚函数后的适应度按下式计算。

$$f_{p,i} = \frac{f_i}{d_i} \quad (6)$$

式中, f_i 是个体 NN_i 的原适应度值, d_i 是个体 NN_i 到它所属类的中心的距离。在施加完罚函数后,种群中最好的 c 个个体被复制到下一代中。

MNN-CREDIT 信用评估模型的具体步骤如下。

步骤 1 初始化种群;

步骤 2 根据适应度函数进行轮盘赌选择操作,并保留当前 c 个最好个体(神经网络);

步骤 3 对种群进行交叉、变异等遗传操作;

步骤 4 把当前种群和上一代中最好的 c 个个体组合成新种群;

步骤 5 对种群执行 K-Means 算法,把种群划分成 c 个类;

步骤 6 除了每个类中最好(潜能值最高)的个体外,对其他个体的适应度值执行罚函数;

步骤 7 根据适应度函数对种群进行选择,并保留当前 c 个最好个体;

步骤 8 如果未达到预先设定的最大进化代数或最好的 c 个个体未达到预先设定的精度要求时,则返回第 3 步,否则转第 9 步;

步骤 9 输出训练好的 c 个个体(神经网络)。

4.3 信用评估过程

在训练好若干个神经网络之后,就确立了信用评估模型。对于待识别数据,按照动态投票法进行分类,即根据参与集成的 c 个神经网络的训练精度进行加权投票,把得票最多的类标作为最终的信用评估结果。具体步骤如下。

步骤 1 将待识别客户分别输入训练好的 c 个神经网络模型;

步骤 2 对训练好的 c 个神经网络,计算每个神经网络的输出;

步骤 3 对每个神经网络的分类及结果,按照动态投票法确定最终的信用评估结果。

5 信用评估的实证研究

5.1 数据来源

采用德国信用数据库数据^[7]验证了本文提出的 MNN-CREDIT 模型的信用评估精度。德国信用数据库包含 1000 份客户信用资料,每个客户包含的 20 个属性,如表 1 所列。

表 1 德国信用数据库中的客户属性

属性名		属性值(离散化)
属性 1	经常账户状况	1:账户余额 < 0 马克;2:0 马克 ≤ 账户余额 < 200 马克;3: ≥ 200 马克;4:无经常账户记录。
属性 2	账户持续时间(月)	\

(下转第 207 页)

- [5] Sagastizabal C A, Solodov M V. Parallel variable distribution for constrained optimization[J]. Computational Optimization and Applications, 2002, 22: 111-131
- [6] Donghoon L, Wiswall M. A parallel implementation of the simplex function minimization Routine[J]. Computational Economics, 2007, 30(2): 171-187
- [7] 陈政洪, 郁松年. 一个基于 QR 分解的并行原-对偶内点算法[J]. 应用科学学报, 2004, 22(4): 549-552
- [8] 骆志刚, 李晓梅. 块三对角线性方程组的一种分布式并行算法[J]. 计算机学报, 2000, 23(10): 1028-1034
- [9] 段治健, 杨永, 马欣荣, 等. 求解带状线性方程组的一种并行算法

- [10] Gondzio J, Sarkissian R. Parallel interior point solver for structure linear programs[J]. Mathematical Programming, 2003, 96(3): 561-584
- [11] Matlab Parallel Computing Toolbox 4 [EB/OL]. http://www.mathworks.com/access/helpdesk/help/pdf_doc/distcomp/distcomp.pdf, 2010-03-24
- [12] Chern M Y, Murata T. A fast algorithm for concurrent LU decomposition and matrix inversions[C]//Proc. 1983 Int'l Conf. on Parallel Processing. New York: IEEE, 1983: 79-86
- [13] 陈国良. 并行算法的设计与分析[M]. 北京: 高等教育出版社, 2008

(上接第 192 页)

属性名	属性值(离散化)
属性 3 贷款历史状况	1: 无贷款记录或所有贷款均按时归还; 2: 在本银行的所有贷款均按时归还; 3: 迄今为止现存贷款按时归还; 4: 过去曾延迟还款; 5: 存在危帐或仍存在贷款(非本银行)
属性 4 贷款用途	1: 新车; 2: 二手车; 3: 家具设备; 4: 收音机或电视机; 5: 家庭用品; 6: 维修; 7: 教育; 8: 度假; 9: 接受再培训; 10: 经商; 11: 其他用途
属性 5 贷款数额	\
属性 6 储蓄存款 账户状况	1: 账户余额 < 100 马克; 2: 100 马克 ≤ 账户余额 < 500 马克; 3: 500 马克 ≤ 账户余额 < 1000 马克; 4: 账户余额 ≥ 1000 马克; 5: 未知或无储蓄存款
属性 7 现工作就业时间	1: 失业; 2: < 1 年; 3: [1 年, 4 年]; 4: [4 年, 7 年]; 5: ≥ 7 年
属性 8 分期付款占月收入百分比	\
属性 9 个人状况 及性别	1: 男性离异或分居; 2: 女性离异、分居或结婚; 3: 男性单身; 4: 男性结婚或同居; 5: 女性单身
属性 10 其他债务或保证金	1: 无; 2: 联合申请人; 3: 保证人
属性 11 现居住状况	\
属性 12 财产状况	1: 有房产不动产; 2: 无房产不动产, 有社保储蓄协议或养老保险; 3: 无房产不动产, 无社保或养老保险, 有汽车或其他(不在属性 6 范围内); 4: 未知或无财产
属性 13 年龄	\
属性 14 其他分期付款计划	1: 银行; 2: 商店; 3: 无
属性 15 房屋状况	1: 租住; 2: 自有; 3: 免费使用
属性 16 在本银行已有存款数目	\
属性 17 工作状况	1: 失业、无技能或非本地居民; 2: 无技能的本地居民; 3: 技术工人或公务员; 4: 经理、自由职业者、高级雇员或官员
属性 18 应抚养人数	\
属性 19 电话	1: 无; 2: 有或已注册
属性 20 是否外籍劳工	1: 是; 2: 否

5.2 训练与测试

本文的 MNN-CREDIT 信用评估模型在实际评估中发现:

1) 学习模型对数据过分特化, 如果使用训练数据导出分类法, 然后信用评估分类法, 可能错误地导致过于乐观的估计。

2) k -折交叉确认是基于给定数据随机选择划分的, 常用评估分类法准确率的技术。

因此, 为了更准确地评估 MNN-CREDIT 模型的准确率, 把 1000 个数据分成 4 个互不相交的子集数据集, 每个子集各 250 个数据。把这 4 个子集按 3:1 的比例组成训练集和测试集, 进行 4-折交叉确认。4 个信用预测结果的平均值与朴

素贝叶斯模型、支持向量机和 k -近邻判别分析方法($k=17$)信用预测结果的精度进行了比较, 如图 5 所示。

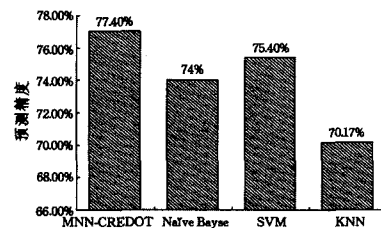


图 5 各信用评估模型的预测精度比较

从图 5 可知, MNN-CREDIT 模型较支持向量机, 其预测精度提高了 1.64%; 较朴素贝叶斯模型, 其预测精度提高了 3.04%; 较 k -近邻判别分析方法($k=17$), 其预测精度提高了 6.87%。

结束语 信用评估对于银行控制风险具有十分重要的意义。传统的信用评估方法存在一些缺陷, 本文针对信用评估对象的特点, 提出了适应信用评估的 MNN-CREDIT 模型, 并利用该模型对德国信用数据库的真实数据进行实证分析。结果表明, MNN-CREDIT 模型较传统的信用评估模型, 取得了较高的预测精度, 证明了本文提出的基于多进化神经网络的信用评估模型的有效性。

参考文献

- [1] 吴江, 唐常杰, 段磊, 等. 基于基因表达式编程的信用评估模型挖掘方法[J]. 计算机应用, 2007, 27(4): 877-880
- [2] 石庆焱, 靳云汇. 个人信用评分的主要模型与方法综述[J]. 统计研究, 2003(8): 36-39
- [3] 王雅林, 赵欣, 黄伟平. KNN 近邻判别分析法在个人信用评估中的应用姜明辉[J]. 数量经济技术经济研究, 2004(2): 143-147
- [4] Krogh A, Vedelsby J. Neural Network Ensembles, Cross Validation, and Active Learning[M]. Advance in Neural Information Processing Systems, 1995, 7: 231-238
- [5] Radolph G. Convergence Analysis of Canonical Genetic Algorithms[J]. IEEE Transactions on Neural Network, 1994, 5(1): 96-101
- [6] Tsymbal A, Pechenizkiy M, Cunningham P. Diversity in Search Strategies for Ensemble Feature Selection[J]. Information Fusion, 2005, 6(1): 83-98
- [7] 赵静娴, 杨宝臣. 一种基于神经网络和决策树的信用评估新方法[J]. 武汉科技大学学报, 2005, 7(2): 36-39
- [8] 张彤, 向南峰. 基于 BP 网络的数字识别方法[J]. 重庆理工大学学报: 自然科学版, 2010, 24(3): 43-47