

文章编号:1001-9081(2009)11-3088-04

## 基于 Logistic 的信用卡套现侦测评分模型

姜 盛

(中国工商银行股份有限公司 牡丹中心,北京 100031)

(j.sheng.1@163.com)

**摘要:**信用卡套现是信用卡产业面临的一种主要风险。单笔的套现交易与正常交易间无显著差异,难以进行基于特征的过滤筛选。为自动、高效地识别套现账户,根据统计学特征先遴选出各相关变量,并结合业务分析,利用 Logistic 回归模型的非线性曲线特征缺陷,克服其自变量多维相关敏感性缺陷,计算出各变量的影响权重系数,构建了信用卡套现侦测评分模型。实践表明识别准确率达到 82.72%。

**关键词:**数据挖掘; Logistic 回归; Logit 变换;  $\chi^2$  检验; 积矩相关系数; 最大似然估计; 马氏距离; K-S 统计量

**中图分类号:** TP311.13; TP39    **文献标志码:**A

## Credit scoring model of detecting illegal cash advance based on Logistic

JIANG Sheng

(Card Center, Industrial and Commercial Bank of China, Beijing 100031, China)

**Abstract:** The illegal-cash-advance is one of the major fraud risks of the credit-card industry. There is almost no difference between single illegal-cash- advance transaction and the normal one, so it could not distinguish them based on different characteristics. To detect the illegal-cash-advance accounts automatically and accurately, the authors picked up correlative variables first, then made a business analysis, and took advantage of the nonlinear curve feature—the one defection of Logistic, and overcame the sensitivity of the multidimensional relativity between independent variables — the another defection, at last computed the weight of coefficient, and constructed a credit scoring model. The applications indicate that the accuracy of the model has achieved 82.72%.

**Key words:** data mining; Logistic regression; Logit transformation;  $\chi^2$  test; product moment correlation coefficient; Maximum Likelihood Estimate (MLE); Mahalanobis distance; K-S statistic

## 0 引言

信用卡套现是指持卡人不通过正常手续提取现金,违反与发卡机构的约定,将信用卡中的信用额度通过 POS 终端或其他方式,全部或部分地直接转换成现金,利用贷记卡的免息还款期优惠政策,而不必向发卡机构支付取款费用及透支利息的行为。信用卡套现是信用卡产业面临的一种主要欺诈风险,我国刑法和相关金融法规将其界定为一种违法行为。单笔的套现交易与正常账户的消费交易间无显著差异特征,不能通过简单的过滤搜索实现套现账户的筛选。本文提出了一种基于 Logistic 回归技术,通过数据挖掘,建立一套信用卡套现侦测模型的技术解决方案,以实现对信用卡套现账户自动、高效、准确的识别。先通过汇总部分已确认的套现账户信息,再取相同数量的正常账户作为对照组,将各种可收集到的指标纳入考量范围,通过其数学特征找出高度相关的变量,从而识别出各种影响套现活动的指标,并通过业务层面的分析对相关变量进行阐释,最后通过 Logistic 回归模型计算出所有相关指标的影响权重系数,并通过数学推导构建出易于实施的评分模型。实践表明,修正后的模型对信用卡套现账户的识别准确率达到 82.72%,完全符合设计目标。采用 Logistic 分类技术,还可灵活调节各指标权重系数以及界定套现的分割点分数,以根据发展态势动态修正模型。

## 1 Logistic 回归选型策略

### 1.1 信用评分模型

数据挖掘<sup>[1]</sup>的任务主要包括:概念描述、关联分析、分

类、预测、聚类和探索式数据分析等。信用评分是指运用先进的数据挖掘技术和统计分析方法,通过对消费者的人口统计学特征、信用历史记录、行为记录和交易记录等大量数据进行系统地分析,挖掘数据中蕴含的行为模式、信用特征,捕捉历史信息和未来信用表现之间的关系,发展出预测性的模型,以一个信用评分来综合评估消费者未来的某种信用表现<sup>[2]</sup>。它是数据挖掘在金融领域的一种特殊应用,主要应用了数据挖掘中的分类和预测技术。

#### 1.2 各种主要技术比较

目前用于构造信用评分模型的分类技术主要有:判别分析法、线性回归模型方法、Logistic 回归模型方法、分类树方法、线性规划方法、神经网络方法和基因算法方法等。

##### 1.2.1 国外在理论层面的比较研究

表 1 单元格的数值是正确分类的百分比,表中行的数据是可比的,列的数据不可比,因为这 5 种研究使用的总体不同,指标定义也不一样。从表 1 中数据可以看出:在每个学者的研究中,各种技术分类精确度的差异并不显著,可解释为分类方法之间相对的相似性,即各种方法基本殊途同归。

##### 1.2.2 国内在实践层面的比较研究

综合分析表 2 中各种评分技术的错误分类率比较结果,可得出以下结论: Logistic 回归无论是在总错误分类率还是在第 1 类错误比率都排名第二,尤其是第 1 类错误比率仅略高于线性规划法。从总错误分类率和第 1 类错误比率综合考查,Logistic 回归方法是一个较好的方法。

收稿日期:2009-05-06;修回日期:2009-07-17。

作者简介:姜盛(1971-),男,山东莱州人,工程师,硕士,主要研究方向:数据挖掘、信息安全。

表 1 不同信用评分技术分类精确度的比较<sup>[3]</sup> %

作者	线性回归	Logistic 回归	分类树	线性规划	神经网络	基因算法
V. Srinivasan	87.5	89.3	93.2	86.1	—	—
M. Boyle	77.5	—	75.0	74.7	—	—
W. E. Henley	43.4	43.3	43.8	—	—	—
M. B. Yobas	68.4	—	62.3	—	62.0	64.5
V. S. Desai	66.5	67.3	—	—	66.4	—

### 1.3 业务需求指标

业务需求可以概要描述如下。

1) 设定一个是否套现的二值变量为因变量。

表 2 各种评分技术的错误分类率比较<sup>[4]</sup> %

方法	训练样本			保留样本		
	总错误分类率	第 1 类错误比率	第 2 类错误比率	总错误分类率	第 1 类错误比率	第 2 类错误比率
判别分析	28.53	27.31	29.74	29.46	25.45	33.48
Logistic 回归	29.14	26.40	31.87	28.79	22.32	35.27
线性规划	30.50	29.14	31.87	30.58	21.43	39.73
神经网络	21.55	25.65	17.40	24.33	25.89	22.76
分类树	21.92	20.64	23.22	28.79	30.36	27.23

## 2 基于 Logistic 的信用卡套现侦测评分模型

### 2.1 Logistic 回归

Logistic 回归主要适用于二元性目标变量,即因变量  $Y$  的值只能是 0 或者 1,1 代表某种结果的发生,而 0 代表某种结果的不发生。其自变量可以是连续性变量,也可以是类别性变量。Logistic 回归的数学模型为:

$$p(y = 1 | \mathbf{X}) = \frac{e^{(\beta_0 + \boldsymbol{\beta}^T \mathbf{X})}}{1 + e^{(\beta_0 + \boldsymbol{\beta}^T \mathbf{X})}}$$

其中: $\mathbf{X}$  是  $m$  维向量; $\boldsymbol{\beta}$  是  $m$  维待估计的参数。

对上式做 Logit 变换<sup>[5]</sup>,Logistic 回归模型可以变换成下列线性形式:

$$\text{logit}(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \boldsymbol{\beta}^T \mathbf{X}$$

Logistic 回归法具有以下优点:1) 预测结果是介于 0 和 1 之间的概率;2) 可以适用于连续性或类别性自变量;3) 容易使用,容易理解。

Logistic 回归的基本思想是利用 Logistic 函数呈 S 型分布、单调上升的特点,通过 Logit 变换,巧妙地将非线性函数  $P$ (事件发生的条件概率)变换成以  $odds = P/(1 - P)$  为自变量的线性函数,以进一步分析、处理。

### 2.2 Logistic 回归技术缺陷

Logistic 回归模型存在两个主要技术缺陷:

1) 预测结果的概率转换呈非线性的 S 型,在两端概率的变化较小,在中间的概率变化很大;

2) 对模型中自变量的多维相关性较为敏感。

针对这两个主要的缺陷,分析应对策略如下:

1) 利用非线性的 S 型转换曲线特性,使得分布在曲线两端的大部分套现账户与正常账户分别有着相近的发生概率,而在中间过渡部分变化很大,可实现两个对立账户状态的快速切换,缩小具有模糊界定特征的区域范围,便于定量划分。

2) 为避免因此特性造成的模型结果失真,需对候选特征

2) 不仅能识别套现账户,还能分离出对其有相关作用的变量,用于对套现的定性描述,因此不宜采用基于黑箱技术的神经网络和基因算法技术。

3) 技术要相对成熟。不宜采用最新的技术,否则还要面临对所使用的新方法在技术可行性、有效性、稳定性等方面的验证。

4) 模型要具有较高的准确性和良好的稳健性。

### 1.4 确定技术选型方案

综合考虑上述因素,决定采用分类准确性较高、性能比较稳健、易理解、适合二元因变量分析的 Logistic 回归技术来构建信用卡套现侦测评分模型。

表 2 各种评分技术的错误分类率比较<sup>[4]</sup> %

变量进行相关性分析,以减少它们之间的相关性,消除变量共线性的负面影响。具体可以利用计算得到的相关系数矩阵,并结合领域知识,以决定剔除那些具备共线性特征,且不具备典型业务特征的变量。

下面从统计学角度,结合领域知识简要介绍剔除高度相关的变量的方法,以避免所构造模型结果失真。

#### 2.2.1 初选特征变量

由于各种具体应用的领域知识背景、商业目的、数据分布、数据质量和数据平台等不尽相同,因此数据挖掘前期数据准备的操作差异性很大。此阶段需要实施人员对业务、数据库、数据结构和含义以及商业目标有全面、深入的理解,在此不再赘述。通过对银行业务系统数据进行综合的业务分析,确定了正常账户及具备套现特征账户各 2 000 条作为训练样本集,并选取等量的数据作为保留样本集。经过整合、转换、去噪和规范化等综合数据处理,初步筛选出部分候选特征变量如表 3 所示(为方便描述问题,所选变量有适度调整和删减)。

表 3 初次确定候选变量清单

序号	变量名	说明	序号	变量名	说明
1	District	地域	10	HeadShip	职务
2	Openmons	有效月数	11	TechTitle	职称
3	CreditClass	信用等级	12	Edu	教育程度
4	Delq	违约数	13	Gender	性别
5	AvgAmtCons	消费均值	14	Occupation	职业
6	AvgAmtCash	取现均值	15	MaxUseCl	额度峰值
7	UseCl	额度使用	16	FstranMon	首交易月
8	ConsAmt_Rate	消费占比	17	Nonsys_Rate	非系统比
9	CashAmt_Rate	取现占比			

#### 2.2.2 检验候选特征变量依赖性

$\chi^2$  检验是利用随机样本对总体分布与某种特定分布拟合程度的检验<sup>[6]</sup>,也就是检验观测值与理论值之间的紧密程度。当研究  $K(K > 2)$  个事件时,可以测定  $K$  个观察值与相

应的理论值之间的差异,为此而构造的统计量称为 $\chi^2$  统计量,亦称 Pearson 定理。该定理说明,当样本容量充分大时,样本分成 $K$ 类,每类实际出现的次数用 $f_0$ 表示,其理论次数为 $f_e$ ,则 $\chi^2$  统计量为:

$$\chi^2 = \sum_{i=1}^k \frac{(f_0 - f_e)^2}{f_e}$$

且服从分布 $\chi^2(K-1)$ ,式中( $K-1$ )为自由度<sup>[7]</sup>。

$\chi^2$  分布原本是一种特定形式的概率分布。在非参数统计中, $\chi^2$  检验常用于判断两组或多组的资料是否彼此关联的问题,如果各组资料彼此不关联,就称为独立,所以这类问题也称为独立性检验。独立性检验的特点是其理论频数不是预先确定的,而需要从样本资料中获得。

下面以特征变量“地域”为例,介绍利用 $\chi^2$  检验来实现独立性检验的过程。

通过数据库操作的聚合函数很容易构造出 Flag 与 District 两个变量进行交叉分类的频数分布表,即列联表,见表 4。

表 4 Flag 与 District 变量构成的列联表

Flag	District1	District2	District3
Flag0	505	612	883
Flag1	1351	526	123

首先提出假设  $H_0$ : 地区类别和套现标志之间是独立的(不存在依赖关系)。接着,通过计算可得到 $\chi^2 = 966.28$ 。 $\chi^2$  的自由度为 $(R-1)(C-1) = 2$ ,取 $\alpha = 0.05$ ,查 $\chi^2$  分布表可知 $\chi^2_{0.05}(2) = 5.991$ ,由于 $\chi^2 > \chi^2_{0.05}(2)$ ,故拒绝原假设,即地区类别和套现标志之间存在依赖关系,套现标志受地区类别影响。

采用同样的方法对其他 16 个变量进行独立性检验,发现它们全部与套现标志之间存在着显著的依赖关系。

### 2.2.3 计算变量间相关系数

积矩相关系数是一测定两变量线性相关的计算方法,计算公式为:

$$r = \frac{\sigma_{xy}^2}{\sigma_x \cdot \sigma_y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \cdot \sqrt{\sum (Y - \bar{Y})^2}}$$

其中: $r$  为积距相关系数; $\sigma_{xy}$  为  $x$  与  $y$  变量的协方差; $\sigma_x$  和  $\sigma_y$  分别是  $x$  和  $y$  变量的标准差。用协方差来测定两变量的线性相关,不仅能直接显示相关的方向,而且可表明两变量的“共变性”。 $r$  是一个无量纲系数,具备以下特征:

- 1) 不受变量值水平和计量单位的影响;
- 2) 相关系数的值有一定的范围,即 $|r| \leq 1$ 。通常的判别标准为: $|r| < 0.3$  表示无相关关系; $0.3 \leq |r| < 0.5$  表示低度相关关系; $0.5 \leq |r| < 0.8$  表示中度相关关系; $0.8 \leq |r| < 1$  表示高度相关关系。

利用 SAS 软件中的 Pearson Correlation 计算功能对包含 1 个因变量和 17 个特征变量的所有变量进行两两相关分析,得到全部 18 个变量两两之间的相关系数矩阵。根据计算结果,可得出以下结论:

- 1) Edu、Occupation 的  $P$  值大于 0.05,表示与因变量之间不存在线性相关关系;
- 2) AvgAmtCash、Headship、TechTitle、Gender、Nonsys\_Rate

的相关系数小于 0.3,表示无线性相关关系;

3) 剩下的 10 个变量都与因变量之间的相关系数 $|r| > 0.3$  且  $p$  值小于 0.05,即呈统计学意义上显著的线性相关关系。

接下来,结合领域知识再对具有统计学意义上显著线性相关的 10 个候选变量进行相关性分析,决定剔除 OpenMons、AvgAmtCons、UseCl、ConsAmt\_Rate 共 4 个变量。至此,候选特征变量只剩 6 个。

### 2.3 计算 Logistic 方程的回归系数

对 Logistic 回归模型的参数估计,通常采用最大似然估计法<sup>[8]</sup>(Maximum Likelihood Estimate, MLE),其统计原理是先对  $n$  例观察样本建立似然函数

$$L(\beta) = \prod_{i=1}^n P_i^{Y_i} (1 - P_i)^{1-Y_i}$$

其中: $P_i = P(Y_i = 1 | X_1, \dots, X_m)$  表示在第  $i$  例观察对象的自变量的作用下阳性结果发生的概率。为简化计算,取似然函数的对数,得到:

$$\ln(L(\beta)) = \sum_{i=1}^n Y_i \ln P_i + (1 - Y_i) \ln(1 - P_i)$$

由于似然函数  $L(\beta)$  与对数似然函数  $\ln(L(\beta))$  有相同的单调性,对  $\ln(L(\beta))$  求导建立对数似然方程,采用 Newton-Raphson 等迭代方法求得参数的估计值  $b_0, b_1, b_2, \dots, b_m$ ,作为 Logistic 回归模型参数  $\beta_0, \beta_1, \beta_2, \dots, \beta_m$  的最大似然估计值。一般当样本总量大于 100 时,Logistic 回归的最大似然估计具有较好的一致性、渐进有效性和渐进正态性。

可以利用 SAS Enterprise Guide 的 Logistic 回归计算功能完成以上繁杂的计算,得到 Logistic 方程的截距及各回归系数。

### 2.4 构造评分模型

将 Logistic 回归方程的截距及各回归系数代入经 Logit 变换后的公式可得:

$$\ln\left(\frac{p}{1-p}\right) = \ln(odds) = \alpha_0 + \alpha_1\beta_1 + \alpha_2\beta_2 + \dots + \alpha_n\beta_n \quad (1)$$

如前所述,  $\ln(odds)$  为 $(-\infty, +\infty)$  内的任一值,并呈线性分布,如要判断某账户的套现概率,需将其各种指标代入式(1),求出  $odds$  值,再通过取对数计算,得出  $p/(1-p)$  的值,最后解一元一次方程,最终得出某账户为套现账户的概率,计算繁琐。为便于推广,需将这些知识以更为直观、便于理解和操作的评分卡形式展现出来,构造评分卡的数学推导过程为:

$$Score = Offset + Factor \cdot \ln(odds) \quad (2)$$

现在设定套现账户与非套现账户的比例相当时分值为 300,并假设当套现账户与非套现账户之比增加一倍时,其分数增加 20 分,代入式(2),可求解  $Factor$  和  $Offset$ 。

将式(1)代入式(2),可得:

$$\begin{aligned} Score &= Offset + Factor \cdot \ln(odds) = \\ &= Offset + Factor \cdot (\alpha_0 + \alpha_1\beta_1 + \alpha_2\beta_2 + \dots + \alpha_n\beta_n) = \\ &= Offset + Factor \cdot \alpha_0 + Factor \cdot \sum_{i=1}^n \alpha_i\beta_i = \\ &= \sum_{i=1}^n \left( \frac{Offset}{n} + \frac{Factor \cdot \alpha_0}{n} + Factor \cdot \alpha_i\beta_i \right) \end{aligned} \quad (3)$$

依照式(3)构造出如表 5 所示的信用卡套现侦测模型评分卡,具体数值用  $S_{mn}$  替代。

某账户得分为所有分量得分的总和,即:

$$Score = S_{1n} + S_{2n} + S_{3n} + S_{4n} + S_{5n} + S_{6n}$$

按照设计目标,如某个账户的得分高于 300 分,应视作具备套现账户特征;如低于 300 分,可视作正常账户。

表 5 信用卡套现侦测模型评分卡

评分项目	分值	评分项目	分值
地区 1	$S_{11}$	取现占比 1	$S_{41}$
地区 2	$S_{12}$	取现占比 2	$S_{42}$
地区 3	$S_{13}$	取现占比 3	$S_{43}$
信用等级 1	$S_{21}$	额度峰值 1	$S_{51}$
信用等级 2	$S_{22}$	额度峰值 2	$S_{52}$
信用等级 3	$S_{23}$	额度峰值 3	$S_{53}$
违约情况 1	$S_{31}$	首交易月 1	$S_{61}$
违约情况 2	$S_{32}$	首交易月 2	$S_{62}$
		首交易月 3	$S_{63}$

### 3 实验结果及分析

需对所构造的 Logistic 回归模型进行三个层面的检验:对模型本身各系数的有效性及可信性进行检验;对训练样本进行模型分类精度的检验;对保留样本进行模型分类精度(也即预测准确度)的检验。

#### 3.1 模型系数检验

Wald 检验为常用的回归系数检验方法,可适用于对单一及整体回归系数的检验。在检验假设为  $H_0: \beta_j = 0$  时,Wald 检验对模型中的单个回归系数的检验公式为:

$$\chi^2_w = \left( \frac{b_j}{S_{b_j}} \right)^2$$

统计量  $\chi^2_w$  渐进服从自由度为 1 的  $\chi^2$  分布。实验所得系数的  $p$  值最大 0.023,最小不足 0.0001,均小于 0.05,说明原假设不成立。进一步解释为:单个变量的系数  $\beta_j$  呈现统计学意义上显著的  $\beta_j \neq 0$  的特征,即所得系数是有效、可信的。

#### 3.2 训练样本分类精度检验

所谓拟合优度检验也就是使用所构建的模型对训练样本进行分类精度的检验。有两个业界常用的衡量指标:定量的马氏距离统计量指标、定性的 K-S 统计量指标。

##### 3.2.1 马氏距离指标检验

马氏距离的定义是两种类别数据的均值之差比上其合并标准差<sup>[3]</sup>,其公式为  $M = \frac{m_x - m_y}{\sigma}$ 。如评分卡的马氏距离越大,则分类器越好。其物理意义是:两个类别之间的距离越大,则表示两类账户(套现、正常)各自的分布越集中,且这两个分布图之间距离越大,分类效果也就越好,如图 1。

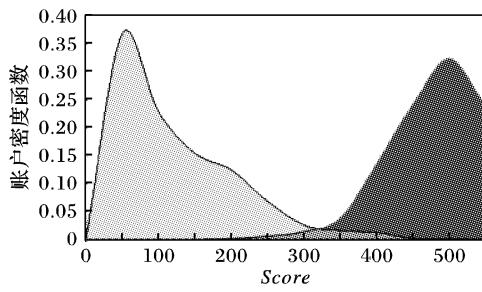


图 1 马氏距离指标示意图

据实验结果求得马氏距离  $M = 4.74$ ,区分能力非常强。

#### 3.2.2 K-S 指标检验

K-S 统计量指标用于度量两种类别的数据累积分布比例函数之间最大的距离。距离越大,K-S 指标越高,模型的区分功能越强<sup>[9]</sup>。

该指标最大的缺点是它只能衡量某个点的背离度,而该点可能不在所期望的分割点附近,尤其是当分割点在评分值域的极值附近时,该指标的缺点表现得尤为明显。此时,较好的解决办法是不衡量两种类别数据的累积分布比例函数之间的最大距离,而是衡量分割点处的两个累积分布比例函数之间的距离,以此作为 K-S 指标的替代值。

以图 2 为例,在 300 分(横坐标)时为预期分界点,此时模型判断其为套现账户的比率与判断为正常账户的比率之差为 94.75%。业界一般认为对行为评分的区分距离 45% 以上就是可以接受的,本方案对训练样本的区分距离达到 94.75%,效果相当理想。而且分割点与预期的设计目标相吻合。

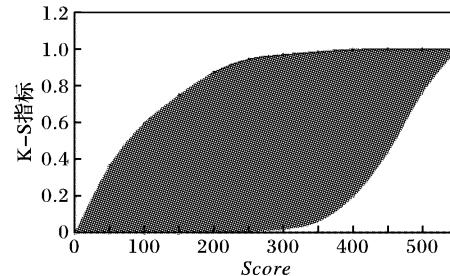


图 2 K-S 指标示意图

#### 3.3 保留样本分类精度检验

保留样本法检验的基本思想是:在建立评分模型时,将样本随机地分成两部分:一部分称为训练样本,用于建立模型;另一部分称为保留样本,用来对模型的分类预测准确度进行检验。

在此,同样利用马氏距离和 K-S 指标来衡量分类精度。根据保留样本,求得马氏距离为  $M = 2.29$ ,区分能力依然非常强。通过计算可以得到,在 300 分这一点上两类数据的距离最大,达到 77.74%,此即 K-S 指标,这表示模型的分类能力非常高,而且分割点与预期的设计目标相吻合。

#### 3.4 模型修正

利用所构建的模型对生产系统中 100 多万条数据进行了评分,抽取全部 500 分以上的账户信息进行检验的结果是:正常账户占比 2.83%,确认套现账户占比 87.10%,具备疑似套现特征账户占比 10.07%。抽取全部 450 分以上、小于等于 500 分的账户信息进行检验的结果是:正常账户占比 13.85%,确认套现账户占比 70.93%,具备疑似套现特征账户占比 15.22%。

考虑到投入成本及具体商业目标,决定修正分界点,由 300 分调整为 450 分,其套现账户识别准确率利用加权平均公式:

$$\frac{N_{450}}{N_{450} + N_{500}} \times \frac{ICA_{450}}{N_{450}} + \frac{N_{500}}{N_{450} + N_{500}} \times \frac{ICA_{500}}{N_{500}}$$

进行计算(为保护商业秘密,具体数值隐去),结果为 82.72%。这个识别准确率完全达到了设计目标的要求。

(下转第 3095 页)

```

if  $R_i$  为空 then 加上一个树叶,标记为训练样本中最普通的类;
else 加上一个由 Generate D-S Decision Tree ( $s_i$ , 候选属性 - 测试属性) 返回的节点;
结束

```

用以上算法处理训练集,所得决策树如图 3 所示。

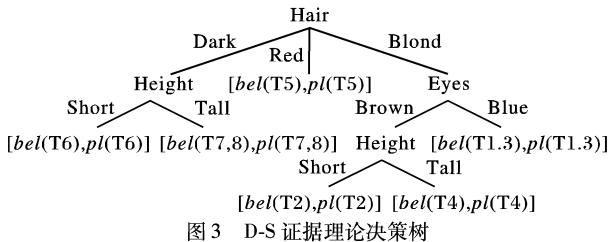


图 3 D-S 证据理论决策树

#### 4 D-S 证据理论决策树的仿真与分析

为测试 D-S 决策树算法的有效性,本文对 D-S 决策树算法进行了计算机仿真。算法用 C 语言实现,运行平台是 Windows XP。训练集为两个,一个是美国加州大学 Irvine 分校(UCI)用于分类算法开发和测试的标准数据集 bollon,从 <http://www.ics.uci.edu/~mlearn/MLRepository.html> 下载。一个是 Quinlan's 的标准实例训练测试集。误差分析采用的是  $K$ -折交差法。标准训练集中的数据都是确定数据,本文采用随机产生  $\Theta$  的一个子集  $\theta$  的方法,把确定数据变成为不确定数据。结果如表 3 所示。

表 3 仿真结果

数据集	所做的测试	属性个数	分类个数	误差率/%
bollon	属性不确定	4	2	15.0
	类不确定	4	2	16.2
golf	属性不确定	4	2	25.0
	类不确定	4	2	14.7

D-S 证据理论决策树算法可以实现对元组所在类或属性不确定的关系数据库进行分类,对大规模的不确定数据进行分类时不会出现组合爆炸问题。D-S 证据理论决策树算法的误差率约为 17.72%。从数据集的分类误差来看,D-S 证据理论决策树和经典的决策树一样,不是对所有的训练集都可以得到很好的效果。本文考虑将 D-S 熵的表达式修正为  $H(R) = \lambda E(R) + (1 - \lambda) N(R)$ ,其中  $\lambda$  的取值范围是 0~1。修正了 D-S 信息熵公式后,通过调节  $\lambda$  的值,可降低分类误差

(上接第 3091 页)

#### 4 结语

在综合分析、研究各种主流的数据挖掘分类技术后,选用 Logistic 回归技术,并利用 Logistic 的非线性特性缺陷、克服自变量多维相关性影响,构建信用卡套现侦测评分模型。对每个主要步骤简要介绍其数学、统计学理论背景或者领域专家的经验知识,结合实际情况设计了适当的解决方案;对所构造的模型进行了全面检验。实验结果表明,所构建的信用卡套现侦测模型具有较好的可靠性和预测分类准确性,分类准确率达到 82.72%,可以自动、高效识别出信用卡套现账户。

#### 参考文献:

- [1] HAN J, KAMBER M. Data mining: Concepts and techniques [M]. San Francisco, CA: Morgan Kaufmann, 2006.
- [2] 陈建. 信用评分模型技术与应用 [M]. 北京: 中国财政经济出版

率。如何快速取得最优的  $\lambda$  值将是下一步的研究内容。

#### 5 结语

本文提出了一个基于证据理论的决策树分类模型。详细论述了在证据理论决策树分类模型中,如何通过证据理论对不确定数据进行表达。给出了用于数据间不确定测量的测量算子和数据间证据合成的聚集算子。将证据理论作为不确定数据表达的有力工具,通过操作算子将证据理论与经典决策树算法相结合,提出了 D-S 证据理论决策树分类算法。用 C 语言编写了 D-S 证据理论决策树分类算法的程序。实验表明本文提出的 D-S 证据理论决策树分类算法能有效地对不确定数据的分类,有较好的分类准确度。经过适当的参数调节,可以对以概率形式表达的不确定数据进行分类,因而比概率决策树更灵活。D-S 证据理论决策树分类算法在不确定测量阶段和证据合成阶段分别使用测量算子、聚集算子,因而不会出现组合爆炸问题,可适用于大规模不确定数据的分类。

#### 参考文献:

- [1] MCCLEAN S, SCOTNEY B, SHAPCOTT M. Aggregation of imprecise and uncertain information in databases [J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(6): 902~912.
- [2] 段新生. 证据理论与决策、人工智能 [M]. 北京: 中国人民出版社, 1993: 13~34.
- [3] 李芳, 韩元杰. 基于证据理论的知识发现分类算法 [J]. 桂林电子工业学院学报, 2004, 24(3): 27~31.
- [4] ANAND S S, BELL D A, HUGHES J G. EDM: A general framework for data mining based on evidence theory [J]. Data & Knowledge Engineering, 1996, 18(3): 189~223.
- [5] LI YU-HE, DAI HAI-HONG. Reducing uncertainties in data mining [C]// APSEC'97 and ICSC'97: Proceedings of the Fourth Asia-Pacific Software Engineering and International Computer Science Conference. Washington, DC: IEEE Computer Society, 1997: 97~105.
- [6] KULASEKERE E C. Representation of evidence from bodies with access to partial knowledge [D]. Coral Gables, Florida, USA: University of Miami, Faculty of Engineering, 2001: 35~49.
- [7] SCOTNEY B, MCCLEAN S. Database aggregation of imprecise and uncertain evidence [J]. Information Sciences, 2003, 155(3/4): 245~263.
- [8] BRYDON M, GEMINO A. Classification trees and decision-analytic feedforward control: A case study from the video game industry [J]. Data Mining and Knowledge Discovery, 2008, 17(2): 317~342.

- [3] THOMAS L C, CROOK J, EDELMAN D. Credit scoring and its applications [M]. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2002.
- [4] 石庆炎, 秦宛顺. 个人信用评分模型及其应用 [M]. 北京: 中国方正出版社, 2006.
- [5] ABRAHAM B, LODOLTER J. Introduction to regression modeling [M]. Pacific Grove, CA: Duxbury Press, 2005.
- [6] BICKET P J, DOKSUM K A. Mathematical statistics: Basic ideas and selected topics, Vol 1 [M]. 北京: 中国统计出版社, 2004.
- [7] REFAAT M. Data preparation for data mining using SAS [M]. San Francisco, CA: Morgan Kaufmann Publishers, 2007.
- [8] LAROSE D T. Data mining methods and models [M]. Hoboken, NJ: Wiley, 2006.
- [9] 陈建. 现代信用卡管理 [M]. 北京: 中国财政经济出版社, 2005.