第25卷第6期 华东交通大学学报 Vol. 25 No. 6
2008年12月 Journal of East China Jiaotong University Dec. ,2008

文章编号:1005-0523(2008)06-0055-04

基于均值聚类的银行客户信用关系分析

许苗村,蒋先刚

(华东交通大学 基础科学学院,江西 南昌 330013)

摘要:针对银行业中客户贷款契约违约风险较高的问题,通过把经济学中的特征分析模型与数据挖掘中的K_MEANS聚类算法相结合,利用现有客户资料,对客户资信评级分类,从而实现对客户信息的高质量管理,降低银行对客户贷款的风险. 实验结果表明,此算法对客户的资信评级具有良好的分类效果.

关 键 词:信用管理;客户资信评级;K_MEANS聚类

中图分类号:TP315 文献标识码:A

对客户的信用状况进行总体评价(即客户资信评级)可以帮助银行对客户信息进行合理的处理. 本文把银行客户资料系统与计算机技术相结合,能够在繁杂而庞大的客户数据中实现对客户级别的高效甄别. 这需要分两步来完成:首先,要对客户的信用进行分析,采用一定的方法来解释客户信息. 最常见的为专家判断法中的5C法,在此方法下,决定性因素是专家的专业技能、主观判断和对某些关键因素的权衡,因此判断结果依赖于专家的经验,本文采用的特征分析模型可以比较客观地对客户信用进行风险评估,并能对评估结果作出解释,便于下一步算法的实现. 其次,要对客户信息进行分类. K_MEANS聚类算法是在现有算法中应用比较广泛的方法,利用K_MEANS聚类来实现对客户进行A(优)、B、(良)、C(差)3个等级的划分,为银行明确客户消费模式,预测客户的风险性,提供了一种有效的衡量方法.

1 客户信息模型的建立

1.1 特征模型分析

特征分析模型是从多年的信用分析经验中发展起来的一种技术方法,它以充分的客户信息为基础,采用特征分析技术对客户财务和非财务因素以及各方面的特征进行区分、描述和归纳分析. 把与信用分析直接相关且与客户信用状况联系最为密切的若干客户特征抽取出来,并对这些不同表现的含义予以说明. 常用的特征分析技术将客户信用信息分为客户特征、优先特征、信用特征三大类特征,每一特征含6个不同项目,三类共包括可获利润、替代能力、付款历史记录、资信证明、资本总额等18个项目. 详见参考文献[1].

特征分析模型在操作时,首先对每一个项目制定一个衡量标准,分为好、中、差3个等级,每个等级对应不同的分值 "好"对应8~10分;"中"对应4~7分;"差"对应0~3分(某个项目信息缺失时,分值取0);其次,根据客户的偏好和经营特点对每一个项目赋予一个权数,18个项目的权数之和为100.

最后,按照四个步骤完成对特征分析分值(特征百分值)的计算:(1) 根据预先制定的评分标准,在1~10范围内,对上述各项指标评分. (2) 用预先给每项指标设定的权数乘以10,计算出每一项指标的最大评分值,再将这些最大评分值相加,得到全部的最大可能值. (3) 用每一项指标的评分乘以该项指标的权数,得出

收稿日期:2008-09-03

作者简介:许苗村(1984-),女,陕西咸阳人,硕士研究生,研究方向为信息处理与软件技术.

每一项的加权评分值,然后将这些加权评分值相加,得到全部加权评分值. (4) 将全部加权评分值与全部最大可能值相比以得出百分比. 百分比越高表示该客户的资信程度越高,越具有交易价值和信用可靠性.

1.2 客户信用分析与解释

表 1 说明了授信人对于客户甲资信状况的基本判断,表 2 给出了特征分析模型加权百分率的信用含义. 详见参考文献[1].

表 1 客户甲特征分析的各参量状态

项 目	客户特征						优先特征						信用特征						总 计
	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	
评价得分	6	5	9	2	4	6	9	5	6	3	4	5	7	7	5	6	9	5	/
权 数	2	4	7	4	5	7	6	4	3	4	9	5	7	4	7	7	9	6	100
加权值	12	20	63	8	20	42	54	20	18	12	36	25	49	28	35	42	81	30	575
最大值	20	40	70	40	50	70	60	40	30	40	90	50	70	40	70	70	90	60	1 000
百分率	56.9%						53.2%						66.3%						57.5%

表 2 特征分析模型加权百分率的信用含义

加权百分率(%)	信用等级	信用含义
0 – 45	C	信用风险不明朗或较高,交易吸引力低,尽量不向其贷款. 如果向其贷款,要严格控制信用额度,时刻监控其信用状况
45.1 – 65	B	信用风险可以接受,具有交易价值,具有发展为长期客户的潜力,可以向其贷款
65.1 以上	A	信用风险小,是吸引力较大的客户,具有良好的长期交易前景,可给予较大信用额度

1.3 客户信息分类方法

K_MEANS 聚类已被广泛应用于数据挖掘中. 它的核心思想是把 n 个向量 $X_j(j=1,2,\cdots,n)$ 分为 c 个组 $G_i(i=1,2,\cdots,c)$,并求每组的聚类中心,使得非相似性(或距离)指标的价值函数(或目标函数)达到最小. 当选择欧几里德距离为组 j 中向量 x_k 与相应聚类中心 c_i 间的非相似性指标时,价值函数可定义为

$$J=\sum_{i=1}^{c}J_i=\sum_{i=1}^{c}\Big(\sum_{k,x_k\in Gi}||x_k-c_i||^2\Big) \tag{1}$$

这里 $J_i=\sum_{k,x_k\in Gi}||x_k-c_i||^2$ 是组 i 内的价值函数. 这样 J_i 的值依赖于 G_i 的几何特性和 c_i 的位置.

如果用一个通用距离函数 $d(x_k,c_i)$ 代替组 i 中的向量 x_k 的非相似性指标,则相应的总价值函数可表示为

$$J=\sum_{i=1}^{c}J_i=\sum_{i=1}^{c}\Big(\sum_{k,x_k\in Gi}d(x_k,c_i)\Big) \tag{2}$$

为简单起见,这里用欧几里德距离作为向量的非相似性指标,本文中的距离计算可用每个客户的客户特征 c_x、优先特征 o_x、信用特征 r_x 的值与 c 个聚类中心的值分别做计算,距离的计算式为

$$\text{Distance}[i]=\sqrt{(c_x-c_i)^2+(o_x-c_i)^2+(r_x-c_i)^2}$$

划分过的组一般用一个 $c\times n$ 的二维隶属矩阵 U 来定义. 如果第 j 个数据点 x_j 属于组 i,则 U 中的元素 u_{ij} 为 1,否则该元素取 0. 一旦确定聚类中心 c_i,可导出如下使式(1)最小的 u_{ij}

$$u_{ij}=\begin{cases}1 & \text{对每个 } k\neq i,\text{如果 } ||x_j-c_i||^2\leqslant||x_j-c_k||^2\\ 0 & \text{其它}\end{cases} \tag{3}$$

如果 c_i 是 x_j 的最近的聚类中心,那么 x_j 属于组 i. 由于一个给定数据只能属于一个组,所以隶属矩阵 U 具有如下性质

$$\sum_{i=1}^{c}u_{ij}=1\quad\forall j=1,\cdots,n \tag{4}$$

且

$$\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}=n \tag{5}$$

另一方面,如果固定 u_{ij}则使(1) 式最小的最佳聚类中心就是组 i 中所有向量的均值

$$c_i = \frac{1}{|G_i|}\sum_{k,x_k\in G_i} x_k \tag{6}$$

这里$|G_i|$是 G_i 的规模或$|G_i| = \sum_{j=1}^{n} u_{ij}$.

对于数据集 x_i,($i=1,2,\cdots,n$) ,一般 K 均值聚类算法用来确定聚类中心 c_i 和隶属矩阵 U 的实现步骤为:

步骤 1: 初始化聚类中心 c_i($i=1,2,\cdots,c$) ,可从所有数据点中随机取 c 个点.

步骤 2: 用式(3) 确定隶属矩阵 U.

步骤 3: 根据式(1) 计算价值函数. 如果它小于某个确定的阈值,或它相对上次价值函数值的改变量小于某个控制值,则获得聚类中心值,算法停止.

步骤 4: 根据式(6) 修正聚类中心,返回步骤 2.

2　客户信用管理系统的算法设计与实验结论

结合特征分析模型及特征分析模型的信用含义可以知道,客户的特征分析模型用加权百分率决定了客户属于哪一类. 由于每一个客户的偏好和经营特点不同,因此实验验证中在客户数据库中选取了 30 万个客户的客户特征、优先特征、信用特征的加权值,并对这些特征值进行归一化处理,用归一化后的特征值代替原始的特征值,可以提高客户信用分类的正确率. 主要方法如下: 全体客户可分为 3 类,每类的客户数为 L_i($i=1,2,3$) . 根据 3 个特征值对客户进行分类. 先求出第 k 种特征值的平均值和方差.

$$\overline{X}_k = \frac{\sum_{i=1}^{3}\sum_{j=1}^{L_i} x_{ijk}}{\sum_{i=1}^{3} L_i} \qquad \sigma^2 = \frac{\sum_{i=1}^{3}\sum_{j=1}^{L_i}\left(x_{ijk}-\overline{X}_k\right)^2}{\sum_{i=1}^{3} L_i} \qquad k=1,2,3 \tag{7}$$

其中,x_{ijk}表示第 i 类第 j 个客户的第 k 种特征值,由此得到归一化的特征值.

$$y_{ijk} = \frac{x_{ijk}-\overline{x}_k}{\sqrt{\sigma^2}} \qquad i=1,2,3,\quad j=1,2,\cdots,L_i,\quad k=1,2,3. \tag{8}$$

为了使用 K_MEANS 聚类算法对客户进行分类,我们事先给定 A、B、C 三个信用等级中客户特征、优先特征、信用特征的加权中心值,如表 3 所示.

表 3　初始聚类中心参量表

	A	B	C
客户特征中心值	x	y	z
优先特征中心值	k	h	g
信用特征中心值	l	m	n
三个等级的中心值	x + k + l = 650	y + h + m = 450	z + g + n = 200

用 K_MEANS 聚类算法按照事先给定的 9 个中心值,让所有客户的三个指标与 A 类、B 类、C 类的三个中心值做相似度度量,与哪一类数据的相似度最大时,就把此客户划分给哪一类.

通过以上算法实现,银行可方便根据已掌握的客户信息对客户进行分类,图 4 显示了初始聚类中心为 463,400,313,140,416,285,288,111,300 和初始聚类中心为 476,344,293,152,434,308,233,120,279 时,对 30 万个客户进行分类,每一个信用等级的客户数量,图 5 则显示了客户甲的信用等级状况,客户甲的加权百分率位于 B 类之上 A 类之下,因此可直接由图表看出客户甲属于 B 类.

由图 4 可以看出,以不同的初始聚类中心进行聚类,得到的结果几乎是完全相同的,误差仅仅只有千分之一,因此对于不同地区的银行来说,设置不同的初始聚类中心,对结果的影响不大,依然可以得到对客户的精确划分,满足了此系统的客户关系管理鲁棒性的要求.

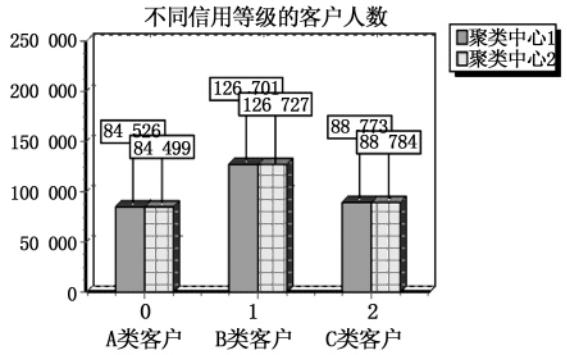

图4 以不同初始聚类中心进行聚类的实验结果

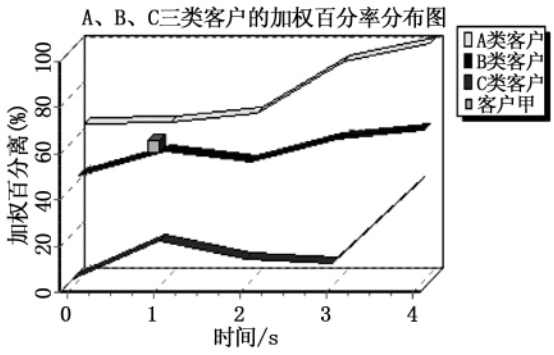

图5 客户甲的信用等级状况

参考文献:

[1] 刘戒骄. 个人信用管理 [M]. 北京: 对外经济贸易大学出版社,2003.

[2] 黄 宪,赵 征,代军勋. 银行管理学 [M]. 武汉: 武汉大学出版社,2004.

[3] 蒋先刚,涂晓斌. 基于模糊均值聚类和图像可控细化技术 [J]. 计算机应用与软件,2008,25(7):86 - 88.

[4] 刘前进,王 蒙,等. Delphi 数据库编程技术 [M]. 北京: 人民邮电出版社,2000.

An Analysis on Credit Relationship of Bank Customer Based on K_MEANS Cluster

XU Miao_cun,JIANG Xian_gang

(School of Basic Sciences,East China Jiaotong University,Nanchang 330013,China)

Abstract: Aiming at the high risk of customer loan contract violation in bank enterprise,combining a characteristic analysis model in economics with K_MEANS cluster algorithm in data mining,this paper carries on customer credit rating by using existing customer materials,thus realizing the high quality management of customer information,reducing the risk of bank loans. The experimental result indicates that this algorithm has good effect upon customer credit rating.

Key words: credit administration; customer credit rating; K_MEANS cluster

(责任编辑: 王建华)