基于数据挖掘的电信客户信用分类模型研究

尤瑞红，李 勇，张博文

（重庆大学 经济与工商管理学院，重庆 400044）

摘要：电信客户信用分析与预测，对于电信运营商在激烈竞争环境下，最大程度的在竞争活动中立于不败之地，具有重要意义。本文采用了SOM聚类算法和用传统经验对客户信用等级评分评级来确定信用等级类标号，再用决策树算法建立两个模型A和B。比较两个模型，选择性能较优者为最终模型并探讨了该模型的实际应用。

关键词：数据挖掘；信用评分；SOM聚类分析；决策树

中图分类号：F270.7 **文献标志码**：A **文章编号**：1671－1807(2010)12－0054－05

电信客户信用分析模型是电信公司在具体管理实践中采取有效措施、减少其信用风险、降低运营成本、提高经济效益的有效途径。而信用评分又是划分客户信用等级的重要技术。

当前，信用评分领域的研究发展速度飞快，为了降低信用决策中的主观因素，越来越多的数学方法被引入到了信用评分中，概括来看，主要分为非统计和统计两大类。统计方法主要包括判别分析、回归分析、分类树和最近邻法，非统计方法包括神经网络、遗传算法、专家系统和数学规划方法。

虽然信用评估专家系统在实践中常有应用，但目前研究这一方法的文章还比较少，并且都不甚深入。这方面的论文主要有Zocco，Davis，Leonard发表的一些文章。[1－2]专家系统解释其信用评估结果的能力很强，这一点对满足一般法律对拒绝申请人贷款申请要给出合理解释的要求有很大帮助。但是有证据显示专家系统对申请人信用变化情况的预测能力很差。Durand[3]将统计学中的判别分析方法引入信用评分模型中，从而在学术界引发了广泛的讨论。代表性的研究有William Fair and Earl Isaacs，Myers and Forgy前者较为完整的采用判别分析法建立了信用评分系统，而后者利用判别分析法对特定领域做了实证分析[4]。近期对于判别分析的研究有Rosenberg，他提出了采用判别分析法进行信用评分可能产生的若干问题[5]。而传统的信用评分模型，回归分析法是目前为止应用最为广泛的，这其中以著名的logistic回归为代表。最早使用回归分析的是Orgle[6]，他采用线性回归模型制定了一个类似于信用卡的评分卡，他的研究表明消费者行为特征比申请表资料更能够预测未来违约可能性的大小。传统的信用决策系统是一个过多依赖于训练有素的专家的主观判断系统，我国电信企业的用户信用管理正处于此阶段，虽具有一定的内容和规范制度，但总体上缺乏科学性、系统性，显然，这样的评分原则主观因素太多，在实际应用中，很难做到客观和真实。

因此，本研究试图利用数据挖掘技术，与电信公司传统的专家决策系统进行建模比较分析，来选择确立哪种方法在管理实践中对电信公司信用等级的分类确定更为有效。最后探讨了该模型的实际应用。

1 研究方法

电信客户信用预测分析是典型的分类预测问题。分类预测分为两个基本步骤：

①以样本数据为训练集(Training Dataset)和测试集(Test Dataset)，以客户信用等级为目标变量建立分类预测模型；②根据分类预测挖掘模型，对客户信用进行分析。数据挖掘技术提供了多种分类预测方法，本文采用了SOM聚类算法和用传统经验对客户信用等级评分评级来确定信用等级类标号，再用决策树算法建立两个模型A和B。比较两个模型，选择性能较优者为最终模型。采用SAS Enterprise Miner和WEKA作为数据挖掘的工具平台。

1.1 数据模型

数据模型是建立客户信用分析模型的前提和条件。数据模型包含目标变量(因变量，由聚类分析得

收稿日期：2010－11－17

作者简介：尤瑞红(1983—)，女，山东临沂人，重庆大学硕士，研究方向：数据挖掘、商务智能、信息系统与决策支持系统、供应链管理；李勇(1969—)，男，四川广元人，重庆大学经济与工商管理学院，副教授，博士，研究方向：研究方向：数据挖掘、商务智能、信息系统与决策支持系统、供应链管理。

到信用等级)以及输入变量(自变量)集合。自变量集合中主要包括:归属地区、客户等级、品牌、性别、年龄、每月的消费金额、每月是否停欠机等变量,我们咨询了该电信运营公司的专家,得出以下数据模型来进行聚类结构如表1。

表1 数据模型

USER_NO	属性名	属性含义	属性类型
1	是否欠停_10	10月份该客户是否欠费停机	binary
2	是否欠停_11	11月份该客户是否欠费停机	binary
3	是否欠停_12	12月份该客户是否欠费停机	binary
4	是否欠停_01	01月份该客户是否欠费停机	binary
5	是否欠停_02	02月份该客户是否欠费停机	binary
6	是否欠停_03	03月份该客户是否欠费停机	binary
7	累计欠费	连续6个月中,该客户账户欠费总额	interval

1.2 SOM聚类

本文采用了SOM聚类,以得到客户的信用等级类标号,由于本文采集的样本数据受到种种局限,故采用的主要是反映内部指标中的簇的凝聚性(SSE)和簇的分离性(SSB)这两个指标来评估。相应公式如下[7]

$$SSE = \sum_{X \in C_i} dist(c_i, X)^2 \quad (1)$$

其中 c_i 表示簇 C_i 的质心

$$SSB = \sum_{i=1}^{K} m_i dist(c_i, c)^2 \quad (2)$$

其中 c_i 表示簇 C_i 的质心;c表示总体质心;m_i 表示簇 C_i 中有 m_i 个个体。

1.3 分类分析

本文就是采用C4.5算法来构建决策树,建立了电信客户信用等级预测模型。决策树可以根据输入变量(自变量)对分类结果影响力的大小,将影响微弱的变量从模型中去掉,从而简化模型。决策树C4.5算法采用信息增益率(gain ratio)作为决策树模型中的属性选择的测试条件,可有效避免传统方法中熵和Gini指标可能产生大量输出的测试条件的情况,提高模型的性能[8]。

设S是s个客户数据样本的集合。根据数据库元组训练集,类标号属性具有m个不同值,因此有m个不同的类 C_i(i=1,……m)。设 s_i 是类 c_i 中的样本数。对一个给定的样本分类所需的期望信息值由下式给出:

$$I = (s_1, s_2, \cdots\cdots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (3)$$

其中 P_i 是任意样本属于 c_i 的概率,并用 S_i/S 估计。

设属性A具有v个不同的值$\{a_1, a_2, \cdots, a_v\}$。可以用属性A将S划分为v个子集$\{s_1, s_2, \cdots\cdots, s_v\}$;其中,$S_j$ 包含S中这样一些样本,它们在A上具有值 a_j。如果A选做测试属性(即最好的分裂属性)则这些子集对应于由包含集合S的节点生长出来的分枝。设 S_{ij} 是子集 S_j 中类 C_i 的样本数。

根据由A划分成子集的熵(entropy)或期望信息为:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j}, \ldots s_{mj}) \quad (4)$$

其项 $\frac{s_{1j} + \ldots + s_{mj}}{s}$ 充当第j个子集的权,并且等于子集(即A值为 a_j)中的样本个数除以S中的样本总数。

熵值越小,子集划分的纯度越高。注意,对于给定的子集 s_j,

$$I(S_{1j}, S_{2j}, \cdots\cdots S_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(P_{ij}) \quad (5)$$

在A上分枝将获得的编码信息(信息增益)是

$$Gain(A) = I(s_1, \ldots, s_m) - E(A) \quad (6)$$

信息增益率(information gain ratio):

$$Gain\ ratio(A) = \quad (7)$$

其中 $Split\ Info(A) = -\sum_{i=1}^{k} p(s_i) \log_2 p(s_i)$,k是属性A将S分成的部分数,$p(s_i)$ 是 s_i 部分占总记录s的比例。

Gain ratio是某个属性导致两个信息量之间的差异率。即具有最高信息增益率的属性,能够最好的反映记录中的不同类的差别,因此选择最高信息增益率的属性作为判定树的测试属性,建立分枝。

2 案例分析

2.1 客户数据

本文所用的所有数据,是从重庆某电信运营商的手机后付费用户数据库中,利用随机抽样的方法提取了从2007年10月至2008年3月的1 101条后付费用户的样本数据,累计属性为37个。该后付费数据

库的样本总量为 8 万条。对初始数据按表 1.1 所示数据结构的要求和电信运营商内部专家的初步处理进行了各种统计和汇总处理，得到 1 101 条样本数据。

2.2　数据预处理

分别在 SAS Enterprise Miner 和 WEKA 两个数据挖掘平台中进行了数据与处理。将各属性变量的中文名修改成其相应拼音缩写，把归属地区统一划分为“主城区”与“非主城”两大类，将性别男女分别用 1 和 0 来表示，年龄划分为 0、1、2 三段。客户等级以 3、2、1 表示，四个品牌以 0、1、2、3 代替。以平均值代替连续变量的缺省值，以众数代替分类变量的缺省值。在 SAS 中以 40%、30%、30%的比例将数据分成训练集(Training Set)、验证集(Validation Set)和测试集(Test Set)。

2.3　模型 B 的建立

2.3.1　SOM 聚类模型输出结果

经过数据的采集、清洗、集成、变换等，SOM 聚类后，输出结果如图 1。

1101	sfqt_10	sfqt_11	sfqt_12	sfqt_01	sfqt_02	sfqt_03	ljqf	_SEGMNT_	Distance	Row	Column	SOM_ID
1	否	否	否	否	否	否	0.0000	1	0.0142	1	1	1:1
2	否	否	否	否	否	否	0.0000	1	0.0142	1	1	1:1
3	否	否	否	否	否	否	0.0000	1	0.0142	1	1	1:1
4	否	是	否	否	否	否	0.0000	2	2.5994	1	2	1:2
5	否	否	否	否	否	否	0.0000	1	0.0142	1	1	1:1
6	否	否	否	否	否	否	0.0000	1	0.0142	1	1	1:1
7	否	否	否	否	否	否	0.0000	1	0.0142	1	1	1:1
8	否	是	是	是	否	是	0.0922	3	2.3247	1	3	1:3

图 1　SOM 聚类结果输出结果

观察图中聚类结果我们可知，原来没有类标号的 1 101 条数据，被自动分成了三个等级，在此结果中多了一列“SEGMENT”，即在非监督条件下对这 1 101条数据进行聚类分析后产生的新的信用等级类标号，我们也可以把他看成分值为 1、2、3 分。

2.3.2　分类分析

根据 2.3.1 节的聚类结果为基础，将每一个电信客户的类标号对应填至“SEGMENT”属性中，然后随机抽取 801 条数据进行分类建模的基本样本数，余下的 300 条数据在模型应用阶段可进行未知类标号的预测，对些来检验该分类模型，既预测出未来电信客户的信用等级。将分类数据导入 SAS 数据库中，运行软件到三叉树的模型。部分三叉树模型如图 2 所示

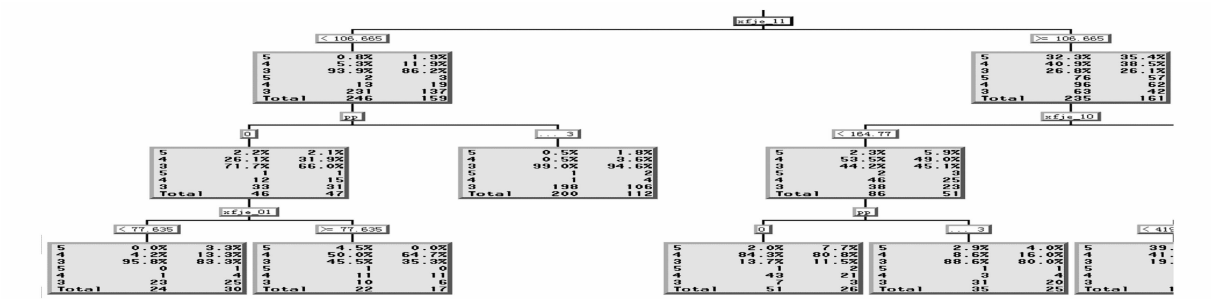

图 2　部分三叉树的树型图

2.4　模型的比较与选择

模型 A 的建立是基于 SOM 聚类技术得到的信用等级新类标号而来的，为了横向比较该模型与之前电信公司仅凭经验而判断的信用等级的好坏进行比较，现将最终分类模型数据中的“SEGMMENT”属性的类标号全部替换成原始类标号(3,4,5)。然后运用 C4.5 算法，依据上述建立模型 A 的方法，建立模型 A(具体步骤省略)

2.4.1　模型的综合评价

在 sas enterprise miner 中应用 assessment node 对模型 A、B 进行评估

1)混淆矩阵。对这两个模型进行评估。通过计算得出模型 B 整体评估正确率为 81.94%，对 3 类电信用户的信用等级的评价指标精度为 90.42%，效果较好。对模型 A，该模型总体正确率不高，只有 56.88%。针对信用分值为 3 的客户正确预测精度也只有 14.29，因此可判断，该模型效果不佳，但基本上可以接受。

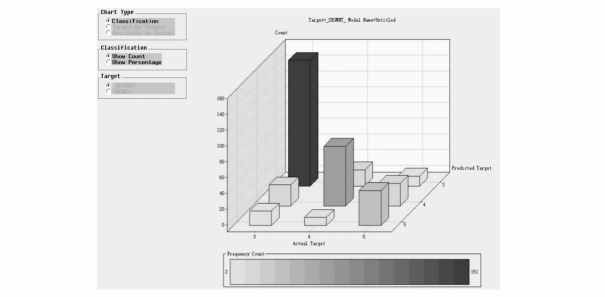

图 3　模型 B 混淆矩阵

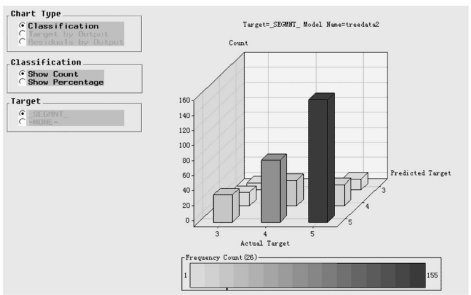

图 4　模型 A 混淆矩阵

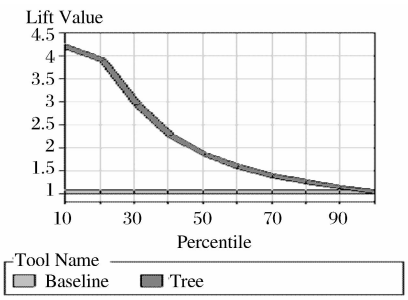

图 5　模型 B lift 图

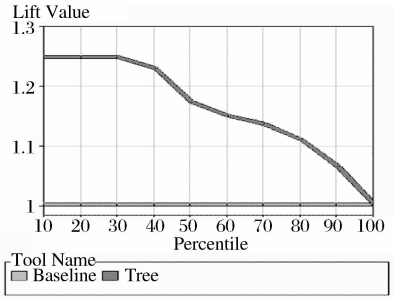

图 6　模型 A lift 图

2)lift 图评估。观察模型 B 的 lift 图(图 5)我们可知，此分类模型前 20%最好的客户的 Lift value 值就达到 3.6 左右，说明 20%左右的客户的预测准确率约为 72%左右，效果较好。观察模型 A(图 6)，分类模型前 30%最好的客户的 Lift value 值仅为 1.25，说明 30%左右的客户的预测准确率约为 37.5%左右。效果明显不如模型 B。

2.4.2　模型的选择

经过一系列的指标计算以后将上述两个分类模型的统计数据汇总至表 2 所示。

通过对该两类模型的评估，可以看出，从模型最重要的整体“正确率”指标可知新模型高出原始模型将近 30%，对可能为信用较低的客户群正确预测的“精度”和“召回率”两个指标，新模型比原始模型要分别高出将近 75%和 82%左右。因此，模型 B 的效果明显好于模型 A。因此我们选择模型 B 即采用先聚类找到客户信用等级的类标号，然后建立三叉决策树模型的方法为最后的客户分类预测模型。

表 2　模型评估统计表

	正确率	错误率	真正率	真负率	假正率	假负率	精度	召回率	最高 Lift 值	Lift 预测准确率
模型 A	56.88%	43.12%	2.56%	64.4%	35.6%	97.44%	14.29%	2.56%	1.25	37.5%
模型 B	81.94%	18.06%	84.36%	73.05%	26.95%	15.64%	90.42%	8.36%	3.6	72%

3　模型的应用

通过数据挖掘技术基于电信客户的基本信息和交易行为数据，建立了客户信用预测模型后，可以将该模型用于预测将来客户的信用问题，尽量降低信用风险，减少呆坏账金额的增加并最终识别出自己感兴趣的知识模式，从而改善本公司的业绩并指导经营管理活动。

3.1　预测客户信用等级类别

利用 SAS Enterprise miner 中的标准 SEMMA 数据挖掘方法，对该电信运营商的客户数据进行探索、分析、建模。找出信用分值最低的那一类，将其标为 3，同理其它两类客户的分值为 4 和 5。在建立该分类模型后即确立电信客户信用等级后，我们可以将其应用到未来的新客户数据，这将减少人力物力成本，提高评价信用等级的自动化程度，进一步提高公司的工作效率。

3.2　营销管理中的应用

根据客户信用等级的差异，有针对性的进行营销管理活动，做到精确营销，这将有效地降低电信运营商的各项成本，并产生良好的效果。我们可以运用该模型采取如下具体措施：根据分类模型所得的分值为 5 的这类客户，信用等级良好，我们可以提高其话费信用额度或者在节假日给予短信问候，寄送小礼物等服务；将营销费用的大额分配到各类信用良好的客户

细分市场，从而有效地做到资源的合理配置；努力将信用分值为 4 的中等信用客户发展成信用较好的一类，积极避免该类客户流失，定时对该类客户电话回访产品、服务质量等措施，可及时把握该类信用客户的动态情况，使得之后的营销工作能够有的放矢。

4 结论

1）使用数据挖掘的各种先进技术，将其运用至电信运营商的信用等级评定过程中，这将提高信用分类的准确力，降低信用风险，增强该电信运营商的综合竞争力。

2）利用 SAS Enterprise Miner 软件，比较分析了基于原始信用等级数据所建立的模型和利用聚类技术确定新的客户信用等级所建立的模型。最终发现新确立的客户信用等级所建立的三叉决策树分类预测性能较好。

3）根据所建立的模型，对该电信运营商的管理工作尤其是在信用管理和营销管理两方面具有极高的应用和指导意义，这将为该运营商针对各类客户提供各种优秀的服务。

参考文献

[1]ZOCCO D P. A framework for expert system s in bank loan management[J]. J. Commercial Bank Lend，1985(67)：47－54.

[2]LEONARD K J. A fraud2alertmodel for credit cards during the authorization process [J]. IMA Journal of Mathematics Applied in Business and Industry，1993(5)：57－ 62.

[3]DURAND D. Risk Elements in Consumer Installment Financing[D]. New York：National Bureau of Economic Research，1941.

[4]李萌 . logic 模型在商业银行信用风险评估中的应用研究[J]. 管理科学，2005 (2).

[5]张维，李玉霜，王春峰 . 递归分类树在信用风险分析中的应用[J]. 系统工程理论与实践，2000(3).

[6]王春峰，万海晖，张维 . 基于神经网络技术的商业银行信用风险评估[J]. 系统工程理论与实践，1999(9).

[7]石庆焱，靳云汇 . 多种个人信用评分模型在中国应用的比较研究[J]. 统计研究，2004(6).

[8]PANG－NING TAN，MICHAEL STEINBACH，VIPIN KUMAR. 数据挖掘导论[M]. 范明，范宏建，等，译 . 北京：人民邮电出版社，2006：2－3.

Research on Telecommunication Customer Credit Classification Based on Date Mining

YOU Rui-hong, LI Yong, ZHANG Bo-wen

(School of Economics and Business Administration，Chongqing University，Chongqing 400044，China)

Abstract: Telecom Customer credit analysis and predicting is of great significance for the telecom operators in a highly competitive environment, the greatest degree of activity in an invincible position in the competition, This paper adopted the SOM clustering algorithm and with traditional experience of customer credit ratings rating to determine the credit rating class label, reoccupy decision tree algorithm of setting two model A and B. To compare the two models, the choice performance is boldness for final model and discusses the model of actual application.

Key words: data mining；credit scoring；SOM cluster analysis；decision tree