# 数据挖掘技术在电信客户流失预警系统中的研究

薛素静，陈 帆
（华北水利水电学院，河南 郑州 450011）

**摘 要**：在明确业务问题的基础上，筛选有效的输入数据和目标变量，并对输入变量各参数之间以及输入变量与目标变量之间进行相关性分析，选取有效的参数。在数据准备完成的基础上，利用 Neural Network 来建立预测模型，并给出预测结果，通过运行实际业务中的数据对模型进行评估。通过该模型预测可能流失的客户，并给出预警信号，以便企业做出经营决策，挽留有关用户，确保企业效益不受影响。

**关键词**：数据挖掘；相关性分析；方差分析；数据建模；神经网络

**中图分类号**：TP311.13 **文献标识码**：A **文章编号**：1674-7720(2011)21-0049-04

## Studies of the data mining technology in the telecom customer churn warning system

Xue Sujing, Chen Fan
(North China University of Water Resources and Electric Power, Zhengzhou 450011, China)

**Abstract**: Based on the clear business problems, the paper screens effective input data and target variables, does correlation analysis between the parameters of the input variables, input variables and target variables, to select effective parameters. On the basis of data, it uses a neural network to establish forecasting model, and gives detection results, evaluates the model by operating the data in the actual business. Through the model to predicate of the loss of customers, and give warning signals, so that the enterprise can make business decisions, keep related users, ensure enterprise efficiency is not affected.

**Key words**: data mining; correlation analysis; variance analysis; data modeling; neural network

客户流失是通信运营商经营中面临的一个基本问题，也是影响经营状况的一个重要因素。一方面，客户离网会造成收入下降、市场占有率下降、营销成本增加、收入降低的问题；另一方面，恶意离网会造成客户恶意欠费，带来不必要的经济损失。新形势下，市场竞争更加激烈，竞争压力与日俱增，迫切需要提升技术支撑能力，为精细化管理和营销等提供有力的信息支撑。

数据挖掘技术是目前数据仓库领域最强大的数据分析手段。它利用已知的数据建立数学模型，找出隐含的业务规则。在客户流失预警分析中，主要方式是根据以前拥有的客户数据，建立客户属性、服务属性和客户消费数据与客户流失可能性关联的数学模型，找出客户属性、服务属性和客户消费数据与客户流失的最终状态的关系，并给出明确的数学公式。市场/销售部门可以根据得到的数学模型随时监控客户流失的可能性。基于严格数学计算的数据挖掘技术能够彻底改变以往电信企业在成功获得客户以后无法监控客户的流失，无法实现客户关怀的状况，把基于科学决策的客户关系管理全面引入到电信企业的市场销售工作中来。

流失预警的目标是通过特定算法分析出哪些客户具有较大的流失概率，从而对这些客户进行有目的、有区别的挽留工作，尽量减少客户流失带来的损失。通过流失模型，提高对高价值客户挽留的成功率，降低客户流失率，降低挽留服务的成本，做到有的放矢，减少由于客户流失带来的收入损失。

## 1 业务理解与数据准备

### 1.1 业务理解

业务问题的定义要求非常明确，任何不明确的定义都会严重影响模型的准确性以及应用时的效果。在客户流失预警分析中，需要明确客户流失的定义。主要有两

个核心的变量：财务原因/非财务原因以及主动流失/被动流失。其中自愿的、非财务原因的流失客户往往是高价值的、稳定的客户。他们会正常的支付自己的服务费用，并对市场活动有所响应。所以这种客户才是电信企业真正想保持的客户。而真正在分析客户流失的状况时，还必须区分公司客户与个人客户，不同服务的贡献率，或者是不同客户消费水平流失标准的不同。举例来说，平均月消费额为 2 000 元左右的客户，当连续几个月消费额降低到 500 元以下时，就可以认为客户发生流失了，而这个流失标准就不能适用于原本平均月消费额就为 500 元左右的客户。实际上，成熟的电信行业客户流失分析经常是根据相对指标判别客户流失。市场调查表明，通常大众的个人通信费用约占总收入的 1%~3%，当客户的个人通信费用降低到远远低于此比例时，就可以认为客户流失发生。所以，客户流失分析系统必须针对各种不同的种类分别定义业务问题，进而分别进行处理。研究发现，客户的流失行为虽然是突发的，但流失前大部分客户原本稳定的话务行为会出现一定程度上的异动，譬如出现交际圈缩小，通话量激剧下降等，本文试图通过分析这种异动来预测客户流失。

### 1.2 数据准备

数据选择包括目标变量的选择、输入变量的选择和建模数据的选择等多个方面。

#### 1.2.1 目标变量的选择

目标变量表示数据挖掘的目标。在客户流失分析中，目标变量通常为客户流失状态。依据业务问题的定义，可以选择一个已知量或多个已知量的明确组合作为目标变量。目标变量的值应该能够直接回答前面定义的业务问题。常规思路一般按照客户流失考核指标（不出账则流失）做目标变量，但这已经处于客户生命周期的晚期，挽留难度大。本研究尝试将时间点提前，从客户价值角度分析客户的生命周期，将目标变量定义在客户价值的急剧下降（Sharp_Decrease）时期[1]。

#### 1.2.2 输入变量的选择

输入变量用于在建模时作为自变量寻找与目标变量之间的关联。在选择输入变量时，通常选择静态数据和动态数据两类数据。静态数据指的是通常不会经常改变的数据，包括服务合同属性（如服务类型、服务时间、交费类型等）以及客户的基本状态（如性别、年龄、收入、婚姻状况、受教育年限/学历、职业、居住地区等）。动态数据指的是经常或定期改变的数据，如每月消费金额、交费纪录、消费特征等[2]。

客户价值定义：本文定义的客户价值主要包含客户通话，因为收入的产生基于客户的消费行为（目前主要考虑是通话），而且从之前的宏观角度看，这些行为更有规律性，具备数据挖掘分析的前提。

流失客户：是指本月有通话，而在之后两个月通话次数小于 15 次，并且在之后两个月通话平均降幅大于 60%的客户。

#### 1.2.3 建模数据的选择

通常电信行业客户流失的方向有两种。第一种是客户的自然消亡，例如由于客户的身故、破产、迁徙、移民等原因，导致客户不再存在；或者是由于客户的升级（如 GSM 升级为 CDMA）造成特定服务的目标客户消失。第二种是客户的转移流失，通常指客户转移到竞争对手享受服务。显然第二种流失的客户才是电信企业真正关心的，是对企业具有挽留价值的客户。因此，在选择建模数据时必须选择第二种流失的客户数据参与建模，才能建立出较精确的模型[3]。

在建立数据模型之前准备以下数据：沉淀本市 2010 年 9 月~12 月的预付费客户和后付费客户的日话单数、月客户资料数据、缴费信息数据以及账单数据。

## 2 建立模型

### 2.1 指标筛选

初步考虑将流失客户分为预付费流失客户及后付费流失客户，沉淀的初始指标包括：

客户个人信息：客户 ID、电话号码、入网日期、主被叫比例、网内外比例、主叫通话时长、被叫通话时长等。

客户通话行为特征：本地通话时常、长途通话时长、总通话时长、短信条数、月租费等。

以上为初始考虑的指标（是实际项目中所列字段的子集），它们当中存在着不同的相关性，并且对预测结果的影响程度也不尽相同，不少指标是噪声指标。为了选出核心指标，通过以下方法进行分析：

（1）指标间的相关性分析[4]。通过该分析选取相互间相关性不高的指标作为考虑指标，这样可以减少指标冗余的现象。

（2）指标与目标结果的相关性分析。通过该分析能够选取出与预测结果相关性高的指标作为考虑指标，这样可选取权重高的指标。

（3）方差分析。通过该分析能够判断出各指标对预测结果的影响程度，从而选取有效的指标。

### 2.2 相关性分析

模型通过计算因素变量间的 Spearman 相关系数来测度变量间的线性相关性。计算过程为：首先把变量值转换为在样本所有变量值中的排列次序，再利用计算方法求解转换后的两个变量对应的排列次序的相关系数，具体计算公式为[5]：

$$r=\frac{\sum(R_{x_i}-\bar{R}_x)(R_{y_i}-\bar{R}_y)}{\sqrt{\sum(R_{x_i}-\bar{R}_x)^2\cdot\sum(R_{y_i}-\bar{R}_y)^2}}$$

其中，$R_{x_i}$ 和 $R_{y_i}$ 分别表示第 $i$ 个 $x$ 变量和 $y$ 变量经过排序后的次序，$\bar{R}_x$ 和 $\bar{R}_y$ 分别表示 $R_{x_i}$ 和 $R_{y_i}$ 的均值。

根据经验，|r|值不同，表示线性相关程度不同：

(1)|r|<0.1 表示弱相关；

(2)0.1<=|r|<0.3 表示低度线性相关；

(3)0.3<=|r|<0.5 表示中低度线性相关；

(4)0.5<=|r|<0.8 表示中度线性相关；

(5)0.8<=|r|<1.0 表示高度线性相关。

对各因素相互间的相关性分析结果如表 1 所示。

表 1 对各因素相互间的相关性分析结果

| 指标 | 当月通话次数 | 零通话天数 | 通话次数均值 | 通话次数标准差 | 通话次数变异系数 | 通话次数比 | 交际客户数 | 对方客服次数 |
|---|---|---|---|---|---|---|---|---|
| 当月通话次数 | 1 | −0.808 34 | 1 | 0.474 14 | −0.690 97 | −0.053 35 | 0.968 53 | −0.002 28 |
| 零通话天数 | −0.808 34 | 1 | −0.808 34 | −0.635 11 | 0.900 89 | 0.064 47 | −0.750 52 | −0.006 86 |
| 通话次数均值 | 1 | −0.808 34 | 1 | 0.474 14 | −0.690 97 | −0.053 35 | 0.968 53 | −0.002 28 |
| 通话次数标准差 | 0.474 14 | −0.635 11 | 0.474 14 | 1 | −0.649 53 | 0.024 2 | 0.388 78 | −0.004 84 |
| 通话次数变异系数 | −0.690 97 | 0.900 89 | −0.690 97 | −0.649 53 | 1 | 0.004 97 | −0.644 17 | −0.004 67 |
| 通话次数比 | −0.053 35 | 0.064 47 | −0.053 35 | 0.024 2 | 0.004 97 | 1 | −0.050 4 | −0.000 66 |
| 交际客户数 | 0.968 53 | −0.750 52 | 0.968 53 | 0.388 78 | −0.644 17 | −0.050 4 | 1 | −0.002 47 |

表 2 是对各因素与目标变量(预测期内通话次数)的相关性分析结果，在模型指标选取时考虑相关性较高的因素作为模型指标。

表 2 对各因素与目标变量的相关性分析结果

| 变量 | 均值 | 标准偏差 | 中位数 | 最小值 | 最大值 | 相关系数 | 结果描述 |
|---|---|---|---|---|---|---|---|
| 本月通话次数 | 106.440 36 | 93.337 94 | 82 | 0 | 385 | 0.834 62 | 高度相关 |
| 通话次数均值 | 3.548 01 | 3.111 27 | 2.733 | 0 | 12.833 | 0.834 62 | 高度相关 |
| 交际客户数 | 77.169 69 | 73.692 83 | 56 | 0 | 358 | 0.825 13 | 高度相关 |
| 忙时通话次数 | 136.059 69 | 183.559 46 | 77 | 0 | 2 629 | 0.733 52 | 中度相关 |
| 本地通话次数 | 125.724 95 | 182.501 54 | 64 | 0 | 2 634 | 0.722 85 | 中度相关 |
| 被叫通话次数 | 77.451 61 | 110.638 46 | 41 | 0 | 1 710 | 0.714 49 | 中度相关 |

### 2.3 方差分析

方差分析是利用样本数据检验待选指标对目标总体影响程度的一种方法。目标总体差异的产生来自两个方面，一方面由总体组间方差造成，即指标的不同水平(值)对结果的影响；另一方面由总体组内方差造成，即指标的同一水平(值)内部随机误差对结果的影响。如果某指标对目标总体结果没有影响，则组内方差与组间方差近似相等；而如果指标对目标总体结果有显著影响，则组间方差大于组内方差。当组间方差与组内方差的比值达到一定程度，或者说达到某个临界点时，就可做出待选指标对结果影响显著的判断[6]。

$$\text{组内方差}=\frac{\sum(x_{ij}-\bar{x}_i)^2}{n_i} \qquad \text{组间方差}=\frac{\sum(\bar{x}_i-\bar{\bar{x}})^2}{n}$$

其中，组间方差表示的是不同套餐所对应的日通话时长之间的差异，$x_{ij}$ 表示日通话时长的第 $i$ 组第 $j$ 个值，$\bar{x}_i$ 表示日通话时长的第 $i$ 组的均值；$n_i$ 表示第 $i$ 组日通话时长的数据个数；$\bar{\bar{x}}$ 表示全体日通话时长的均值；$n$ 表示全体日通话时长分组个数。表 3 是对选取指标方差分析的结果。

表 3 对选取指标方差分析的结果

| 序号 | 指标 | 组间方差 | 组内方差 | 方差比值 | 对目标变量影响程度 |
|---|---|---|---|---|---|
| 1 | 零通话天数 | 12 279 098.6 | 13 620.8 | 901.5 | 非常显著 |
| 2 | 付费次数 | 14 582 512.5 | 28 238.1 | 516.41 | 非常显著 |
| 3 | 付费金额 | 6 361 630.6 | 26 186.8 | 242.93 | 非常显著 |
| 4 | 交际客户数 | 1 265 146.2 | 10 103.1 | 125.22 | 非常显著 |
| 5 | 本月通话次数 | 1 179 191.7 | 9 762.2 | 120.79 | 非常显著 |

表 4 是在对各指标与目标变量之间的相关性及对目标变量结果的影响程度分析的结果综合考虑后，得出的关键指标分析结果。此表为简表，省略了部分内容。

### 2.4 建立模型

采用 Neural Network 来建立预测模型。要预测的对象是截止到当前月月底状态为“正常在用”的客户，预测的目标客户是在本月通话而在未来 60 天通话次数小于 15 次且平均通话降幅大于 60% 的客户，这一结果作为预测模型的输出。神经网络可以从一组输入数据中进行学习，根据这一新的认知调整模型参数，以发现数据中的模式。

表 4 关键指标分析结果

| 序号 | 指标 | 目标影响程度 | 选取程度 |
|---|---|---|---|
| 1 | 零通话天数 | 非常显著 | 重点选取 |
| 2 | 交际客户数 | 非常显著 | 重点选取 |
| 3 | 通话次数均值 | 非常显著 | 重点选取 |
| 4 | 通话次数变异系数 | 较显著 | 重点选取 |
| 5 | 充值金额 | 非常显著 | 重点选取 |
| 6 | 付费次数 | 非常显著 | 重点关注 |

用于训练神经网络的样本数据为数据集中随机抽取样本，各分成三组分别生成网络，比对并验证模型的稳定性。

$$\text{命中率}=\frac{\text{预测命中流失客户数}}{\text{预测流失客户数}}$$

$$\text{覆盖率}=\frac{\text{预测命中流失客户数}}{\text{实际流失客户数}}$$

表 5 中的内容是对本月正常通话，而未来 60 天流失的预付费客户的预测及探索结果。其中，指标集[1]为零通话天数，[2]为交际客户数，[3]为通话次数均值，[4]为通话次数变异系数，[5]为充值金额，[6]为缴费次数，[7]为在网时长。

以通话行为为基础衍生出最终建模所需的变量，并对衍生变量与流失目标进行相关性分析。

(1)零值通话天数与未来通话次数成反比关系，即近期通话越稀疏，不久的将来流失概率越高。

(2)通话次数均值与未来通话次数成严格的正比关系，主动呼叫越频繁，客户越不容易流失。

(3)通话次数标准差与未来通话次数成非线性关系。

表 5 在本月正常通话，而未来 60 天流失的预付其客户的预测

| 试验序号 | 样本数 | 所选指标 | 命中率 | 覆盖率 |
|---|---|---|---|---|
| 1 | 10000 | [1][2][3][4][5][6][7][8][9][10][11][12][13][14] | 51.1% | 53.6% |
| 2 | 10000 | [1][2][3][4][5][6][7][8][9][10][11][12][13] | 53.5% | 55.8% |
| 3 | 10000 | [1][2][3][4][5][6][8] | 54.4% | 51.5% |
| 4 | 10000 | [1][2][3][4][5][6][7] | 50.7% | 57.3% |
| 5 | 10000 | [1][2][3][4][5][6][7][8][9][10][12][13][14][15] | 53.5% | 51.6% |
| 6 | 10000 | [1][2][3][4][5][6][7][8][9][10] | 49.8% | 53.4% |
| 7 | 10000 | [1][2][3][4][5][6][7][8][9][10][11][12] | 50.2% | 52.1% |
| 8 | 10000 | [1][2][3][4][5][6][7][8][9] | 48.9% | 50.4% |

(4) 通话次数的变异系数与未来通话次数成强烈的反比关系，说明近期波动幅度变大是流失的征兆之一。

(5)交际客户数与未来通话次数成强烈的对比关系，说明交际客户数越大客户越不容易流失。

(6)无论客户近期通话突出现明显的增或者突减，客户的流失概率都会增加。

## 3 模型评估

模型的评估应该利用未参与建模的数据进行，这样才能得到准确的结果。检验的方法是对已知客户状态的数据利用模型进行预测，得到模型的预测值，再与实际的客户状态相比较。预测正确值最多的模型就是最优的模型。

表 6 是对 2010 年 10 月份~2011 年 1 月各月正常通话，之后 60 天通话次数小于 15 次且通话平均降幅大于 60%的后付费客户(流失客户)的预测及验证结果。

模型建立后，通过运行实际业务中的数据对模型进行评估，该模型是有效的，实现了客户流失的提前预警，分析了客户流失的可能性，以便运营商在客户流失之前采取有效措施，挽留客户，避免企业的利润受损，实现了系统开发的目标。

表 6 后付费客户流失预测

| | 本月通话客户数 | 之后 60 天流失客户数 | 打分 | 抽取实际 | | | |
|---|---|---|---|---|---|---|---|
| | | | | 抽取客户数 | 流失客户数 | 命中率 | 覆盖率 |
| 2010 年 10 月 | 234502 | 19329 | >-0.3 | 23559 | 11144 | 0.47303 | 0.57654 |
| 2010 年 11 月 | 266850 | 23438 | >-0.3 | 29838 | 12314 | 0.4127 | 0.52539 |
| 2010 年 12 月 | 305920 | 38747 | >-0.3 | 45981 | 25417 | 0.55277 | 0.65597 |
| 2011 年 1 月 | 319552 | 33653 | >-0.3 | 42549 | 22708 | 0.53369 | 0.67477 |


参考文献

[1] 陈志泊.数据仓库与数据挖掘[M].北京：清华大学出版社，2009.

[2] 纪希禹.数据挖掘技术应用实例[M].北京：机械工业出版社，2009.

[3] 杨莉萍，杨晓红.Office Web 组件在 OLAP 分析系统中的应用[J].计算机系统应用，2004(11)：70-72.

[4] 严任远.基于数据仓库的企业 OLAP 多维模型的设计与实现[J].情报杂志，2006，25(9)：33-31.

[5] TREMBA J，LIN T Y. Attibute transformation for data mining：applications to economic and stock market data[J]. International Journal of Intelligent Systems，2002，17(2)：223-223.

[6] WRIGHT R G，KIRKLAND L V，CICCHIANL J，et al. Maintenance data mining and visualization for fault trend analysis [J]. AUTOTESTCON Proceedings，IEEE Systems Readiness Technology Conference，2001：808-815.


(收稿日期：2011-06-09)


作者简介：

薛素静，女，1969 年生，硕士，副教授，主要研究方向：数据仓库、数据挖掘。

陈帆，女，1970 年生，本科，高级工程师，主要研究方向：数据处理。