

硕士研究生学位论文

|  |  |  |
| --- | --- | --- |
| 题目： | | 电信用户流失预警与个性化 |
|  | 推荐的集成研究与应用 | |

|  |  |
| --- | --- |
| 姓 名： | 戴威 |
| 学 号： | 1401210919 |
| 院 系： | 软件与微电子学院 |
| 专 业： | 软件工程 |
| 研究方向： | 物联网技术与应用 |
| 导师姓名： | 李杰 教授 |

二〇一六年七月

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

# 摘要

随着信息技术的发展，电信企业之间的竞争愈加激烈，电信运营商不断推出新的套餐和新的业务，希望能够争取到更多的用户以扩展市场份额，但同时也在很大程度上增加了用户的不稳定性，使得用户流失现象频繁发生。当前学术界虽然对用户的流失预警进行了比较深入的研究，但是主要集中在了流失用户的识别上，对流失用户的维系挽留也没有给出一套行之有效的解决方案，而且在实际应用中模型的准确性和稳定性也有待进一步的提高。

本文的研究目标旨在建立一套从流失用户的识别到维系挽留的解决方案。以联通A省分公司的实际数据为基础，选取31656条在2015年11月具有正常消费行为的用户记录数据，通过采用两步聚类算法构造具有较大差异性的训练样本子集，同时将所构造的训练样本子集分别应用于Logistic回归、决策树、BP神经网络和SVM，构造基分类器，然后将基分类器采用多数投票法进行组合从而形成了新的用户流失预警组合模型。同时本文选取在2015年12月转4G的38294个用户，采用KNN算法建立个性化4G套餐推荐系统，从而对流失用户的维系挽留提供了一种新的策略。

本文通过实际数据比较了在用户流失预警模型中经常采用的单一分类器和本文所构建的组合模型在准确率、精确率、召回率和稳定性上的性能差异，数据表明本文所构建的用户流失预警组合模型的准确率达到了74.8%，优于单一采用Logistic回归、C5.0、CART、BP神经网络和支持向量机所建立的用户流失预警模型。同时对个性化4G套餐推荐系统的离线测试也表明在测试集上的准确率达到了61.66%，取得了满意的结果。

关键词：流失预警 个性化推荐 决策树 Logistic回归 BP神经网络 SVM 组合模型

Integrated Research and Application of Churn Prediction and Personalized Recommendation of Telecom users

Dai Wei( Software Engineer)

Directed by ProfessorJie Li

ABSTRACT

With the development of information technology, the competition among the telecom enterprises is becoming more and more fiercely. Telecom operators are constantly introducing new packages and new business, hoping to gain more customers to expand market share. But also to large extent increase the user's instability, causing the customers churn frequently. Although the current academic circles study on loss warning to users in-depth, which mainly concentrated in the loss of the user's recognition, nor given a set of effective solutions for the bleeding users’ retention.And the accuracy and stability of the model should be improved in practical application.

The aim of this paper is to establish a set of solutions from the identification of the churn to maintain retention. Take a province of the company's actual data of China Unicom as the basis. The two-step clustering algorithm is used to construct the training sample subset which is consisted of 31656 normal users’ recording data in November 2015 with large difference. At the same time, constructing the base classifier using logistic regression, decision tree, BP neural network and SVM.Then the base classifier is combined with the majority voting method to form a new model of user churn prediction. Meanwhile this paper selects 38294 users to 4G in December 2015, using the KNN algorithm to establish the personalized recommendation system of 4G package, providing a new strategy for the retention of the loss of users.

The paper compared accuracy, precision, recall rate and stability performance differences in customer churn early-warning model which is often used in single classifier and the construction of combination model through the actual data. The result showed that the construction of churn predicting model has the best performance74.8% which is superior to the single use of Logistic regression, C5.0, CART, BP neural network and support vector machine established by the customer churn warning model.At the same time, it also shows that the system has achieved quite high accuracy with 61.66% in the offline test of personalized 4G package recommendation system.

KEY WORDS: Churn Prediction，Personalized recommendation, Decision tree, Logistic regression, BP neural network, SVM combination model

# 目录

[第一章 绪论 1](#_Toc448278514)

[1.1 选题背景及研究意义 1](#_Toc448278515)

[1.2 国内外研究现状 2](#_Toc448278516)

[1.3 技术路线与内容结构 5](#_Toc448278517)

[1.3.1 技术路线 5](#_Toc448278518)

[1.3.2 内容结构 6](#_Toc448278519)

[第二章 用户流失预警与个性化推荐的相关理论基础 8](#_Toc448278520)

[2.1 用户流失预警建模所采用的技术 8](#_Toc448278521)

[2.2 组合模型概述 10](#_Toc448278522)

[2.2.1 组合分类器的优势 10](#_Toc448278523)

[2.2.2 基分类器的选择及组合 11](#_Toc448278524)

[2.3 基于4G套餐的个性化推荐方法 13](#_Toc448278525)

[2.4 模型评估 14](#_Toc448278526)

[第三章 用户流失预警单一模型的分析与比较 18](#_Toc448278527)

[3.1 用户数据的选择 18](#_Toc448278528)

[3.2 异常值的处理 25](#_Toc448278529)

[3.3 logistic 回归模型 28](#_Toc448278530)

[3.3.1 logistic回归模型的原理 28](#_Toc448278531)

[3.3.2 logistic回归在用户流失预警模型中的测试 30](#_Toc448278532)

[3.4 决策树模型 33](#_Toc448278533)

[3.4.1 C5.0模型的原理 34](#_Toc448278534)

[3.4.2 用户数据的离散化 35](#_Toc448278535)

[3.4.3 CART模型的原理 37](#_Toc448278536)

[3.4.4 决策树在用户流失预警模型中的测试 38](#_Toc448278537)

[3.5 BP神经网络模型 41](#_Toc448278538)

[3.5.1 BP神经网络模型的原理 41](#_Toc448278539)

[3.5.2 BP神经网络在用户流失预警模型中的测试 44](#_Toc448278540)

[3.6 支持向量机模型 47](#_Toc448278541)

[3.6.1 支持向量机模型的原理 47](#_Toc448278542)

[3.6.2 支持向量机在用户流失预警模型中的测试 51](#_Toc448278543)

[3.7 流失预警单一模型的比较 51](#_Toc448278544)

[第四章 流失预警组合模型的构建与测试 53](#_Toc448278545)

[4.1 训练样本子集的选取 53](#_Toc448278546)

[4.1.1 两步聚类 53](#_Toc448278547)

[4.1.2 样本子集选取的测试 56](#_Toc448278548)

[4.2 用户流失预警组合模型的构建 57](#_Toc448278549)

[4.3 用户流失预警组合模型的测试 58](#_Toc448278550)

[4.3.1 基分类器差异性的测试 58](#_Toc448278551)

[4.3.2 用户流失预警组合模型准确率的测试 59](#_Toc448278552)

[4.3.3 用户流失预警组合模型稳定性的测试 61](#_Toc448278553)

[第五章 个性化4G套餐推荐系统的设计与测试 63](#_Toc448278554)

[5.1 个性化推荐在4G套餐上的应用 63](#_Toc448278555)

[5.2 基于KNN算法的个性化推荐 65](#_Toc448278556)

[5.3 个性化4G套餐推荐系统的测试 66](#_Toc448278557)

[第六章 总结 72](#_Toc448278558)

[参考文献 74](#_Toc448278559)

[致谢 77](#_Toc448278560)

[北京大学学位论文原创性声明和使用授权说明 78](#_Toc448278561)

图表目录

[图1.1 中国联通2015年用户保有量 2](#_Toc448263298)

[图1.2 本文技术路线 6](#_Toc448263299)

[图2.1 ROC曲线 17](#_Toc448263300)

[图3.1 输入变量与目标变量相关性检验 23](#_Toc448263301)

[图3.2 分类变量频数检验 24](#_Toc448263302)

[图3.3 logistic回归变量选择迭代结果 31](#_Toc448263303)

[图3.4 C5.0算法建立的流失预警模型规则集 39](#_Toc448263304)

[图3.5 CART算法建立的流失预警模型规则集 39](#_Toc448263305)

[图3.6 C5.0决策树预测变量重要性排序 40](#_Toc448263306)

[图3.7 单隐层BP神经网络结构图 42](#_Toc448263307)

[图3.8 BP神经网络算法流程图 44](#_Toc448263308)

[图3.9 采用BP神经网络建立流失用户模型的结构 46](#_Toc448263309)

[图3.10 支持向量机的最优超平面 47](#_Toc448263310)

[图4.1 采用两步聚类的聚类结果 57](#_Toc448263311)

[图4.2 流失预警组合模型的构建流程 58](#_Toc448263312)

[图5.1 使用KNN算法寻找最近邻集示意图 65](#_Toc448263313)

[图5.2 属性变量的重要性检测结果 68](#_Toc448263314)

[图5.3 K值和推荐准确率的关系图 69](#_Toc448263315)

[表1.1 2013年至2015年三大运营商用户数量变化表 1](#_Toc448263316)

[表2.1 两个基分类器分类结果 11](#_Toc448263317)

[表2.2 基分类器组合分类结果表 13](#_Toc448263318)

[表2.3 混淆矩阵的结构 16](#_Toc448263319)

[表3.1 初始字段信息提取表 18](#_Toc448263320)

[续表3.1 初始字段信息提取表 19](#_Toc448263321)

[续表3.1 初始字段信息提取表 20](#_Toc448263322)

[表3.2 流失预警模型所选变量 24](#_Toc448263323)

[续表3.2 流失预警模型所选变量 25](#_Toc448263324)

[表3.3 数值变量的异常值检验 27](#_Toc448263325)

[续表3.3 数值变量的异常值检验 28](#_Toc448263326)

[表3.4 logistic回归效果评价的混淆矩阵 33](#_Toc448263327)

[表3.5 连续变量的分箱结果 36](#_Toc448263328)

[续表3.5 连续变量的分箱结果 37](#_Toc448263329)

[表3.6 C5.0决策树模型效果评价的重合矩阵 40](#_Toc448263330)

[表3.7 CART决策树模型效果评价的混淆矩阵 41](#_Toc448263331)

[表3.8 BP神经网络效果评价的混淆矩阵 46](#_Toc448263332)

[表3.9 支持向量机效果评价的混淆矩阵 51](#_Toc448263333)

[表3.10 流失预警模型效果比较 52](#_Toc448263334)

[表4.1 C5.0算法应用于不同的训练样本子集的差异性测试表 59](#_Toc448263335)

[表4.2 不同算法应用于同一训练样本子集的差异性测试表 59](#_Toc448263336)

[表4.3 流失预警组合模型的测试结果表 60](#_Toc448263337)

[表4.4 流失预警单一模型与组合模型的测试结果比较 60](#_Toc448263338)

[续表4.4 流失预警单一模型与组合模型的测试结果比较 61](#_Toc448263339)

[表4.5 组合模型与单一模型稳定度测试对比表 61](#_Toc448263340)

[续表4.5 组合模型与单一模型稳定度测试对比表 62](#_Toc448263341)

[表5.1 用户消费属性字段选择 66](#_Toc448263342)

[表5.2 用户4G套餐迁转数据 67](#_Toc448263343)

[表5.3 K值选择测试结果表 68](#_Toc448263344)

[续表5.3 K值选择测试结果表 69](#_Toc448263345)

[表5.4 KNN混淆矩阵 70](#_Toc448263346)

# 第一章 绪论

## 1.1 选题背景及研究意义

客户流失是许多行业关注的一个重要问题[1]，随着电信行业的迅速发展，各国电信业市场的竞争日益激烈 [2-13]。在我国，激烈竞争的电信市场促使三大电信运营商不断推出新的套餐和新的业务，希望能够争取到更多的用户以扩展市场份额，增加主营业务收入。由于电信市场日趋饱和，获取新用户的成本比留住现有用户要昂贵很多，开发一个新用户的费用大约是维持一个老用户的 5～6倍[14]，而挽留客户的成功率却是发展新客户成功率的16倍[15] ，在电信行业，每年欧洲电信业客户流失率为25%，美国为37%，而在亚洲高达48%[16]，中国移动的客户流失率为12.6%[17]。Reichheld等人[18]的研究表明，客户流失率减少5%，能给企业带来30%～85%的利润增长。由此可见，用户流失将会对运营商的业绩造成巨大的影响。用户流失给运营商带来的主要问题有：第一，主营业务收入下降；第二，用户欠费率增加，并且坏账率也急剧增加；第三，营销成本将会增加。

客户流失分析作为运营商经营分析系统中的一个重要内容，其主要的任务就是根据用户的基本属性和消费行为，建立用户流失预警模型，从而区别流失用户和正常用户。

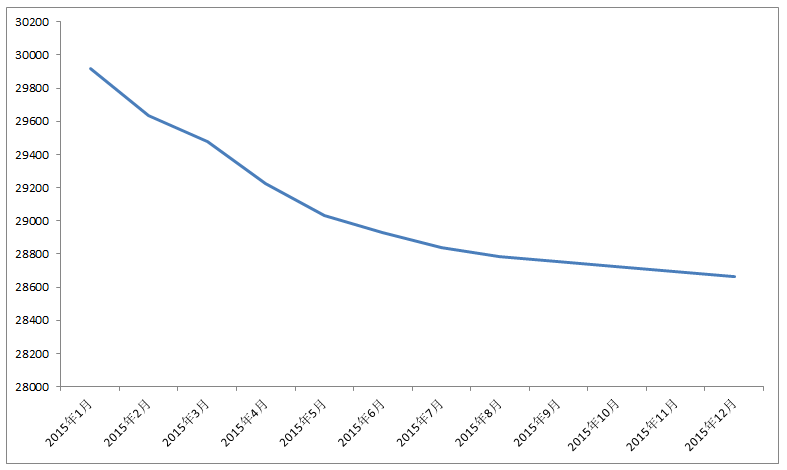
通信信息网的数据显示，三大运营商在4G用户的发展和存量用户的维持上出现了明显的差异，三大运营商从2013年到2015年的用户数量如表1.1所示：

#### 表1.1 2013年至2015年三大运营商用户数量变化表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 运营商 | 统计指标 | 2013年 | 2014年 | 2015年 |
| 中国移动 | 移动电话用户（亿） | 7.67 | 8.06 | 8.26 |
| 4G用户（亿） | 未统计 | 0.9 | 3.12 |
| 3G用户（亿） | 1.92 | 2.46 | 1.69 |
| 中国电信 | 移动电话用户（亿） | 1.86 | 1.86 | 1.98 |
| 3/4G用户（亿） | 1 | 1.2 | 1.4 |
| 中国联通 | 移动电话用户（亿） | 2.81 | 2.99 | 2.87 |
| 3/4G用户（亿） | 1.23 | 1.49 | 1.84 |
| GSM用户（亿） | 1.58 | 1.5 | 1.03 |

表中数据表明，中国移动在发展4G用户方面远远走在了中国电信和中国联通前面。三大运营商在2015年总的移动电话用户总数为13.11亿户，和2014年相比，移动电话用户总数只增加了1.5%，数据充分说明，当前我国的移动电话用户数量已经基本饱和，运营商发展新用户来源于两个方面：一是发展异网用户；二是发展未入网用户，发展异网用户涉及到很多方面，包括网络质量、营销活动等，这将会造成运营商成本的急剧增加，这种方式需要从长期并且结合运营商的发展战略上加以考虑。由于电信市场的饱和，发展未入网用户的难度也非常大。因此在当前，对运营商而言，用户的维系挽留就处于一个非常突出的地位。

2015年全年中国联通用户的保有量的变化如图1.1。



##### 图1.1 中国联通2015年用户保有量

中国联通2015年用户保有量的数据显示，联通用户的数量呈现出明显的下降趋势，因此存量用户的维系对中国联通而言显得尤为急切。

本文旨在通过分析现有的建立在单一分类算法基础之上的用户流失预警模型的不足的基础上，以联通A省分公司BSS平台上现有数据为基础，对包含用户各种属性变量的数据如用户基本属性、话单数据等进行分析，从中提取出关键字段，采用组合分类算法的方式构建新的用户流失预警组合模型，同时根据用户的历史消费记录构建流失用户个性化4G套餐推荐系统。从而形成一套流失用户从识别到维系挽留的策略。

## 1.2 国内外研究现状

客户流失预警的研究始于20世纪90年代。21世纪初，我国学者也开始进行客户流失预测方法的研究，经过十多年的发展，客户流失预测取得了很大进步。通过对目前客户流失预警方法的研究发现，目前客户流失预警使用的主要技术理论是分类算法。具体来说，可以分成两类：一类是基于传统统计学的分类方法；另一类是是人工智能方法。在客户流失预警中使用的统计方法主要有：logistic回归、决策树、贝叶斯分类算法、聚类算法等。这类方法所具有的特点主要是可以对类别型变量和连续型变量进行处理，并且所构建的模型具有很好的解释性。夏国恩[19]首次将KPCA引入到客户流失预警中，提出了基于KPCA的客户流失特征选取方法，设计了基于KPCA的Logistic预测模型，从而实现了对客户流失进行预测，并取得了较好的结果。邹竞和谢鲲[20]使用湖南某移动通信企业2003年后入网的客户数据采用C4.5算法建立了客户流失预警模型，实现了对客户的分类和预测。陈捷[21]使用Clementine12.0，严格按照CRISP-DM数据挖掘标准流程，采用C4.5算法对客户流失进行分析，获得了较好的准确率。郝梅[22]建立了基于CART算法的客户流失预测模型，实验数据表明该模型具有较高的准确率。邓全[23]使用C5.0和CART分别建立客户流失模型，数据表明C5.0算法的覆盖率较高，而CART的准确率较高。赵冬梅和刘贵全[24]提出了一种基于时序模式匹配的k-近邻分类方法，并结合了实际数据和C5.0进行了比较分析，表明该模型具有更好的预测准确率。Wei chiping（2002）[25]等人在假设数据类别是对称的前提下，以客户的合同信息和通话行为作为历史数据，采用决策树C4.5算法来预测客户流失，获得了较高的提升率。朱世武等人[26]利用决策树CHAID算法，对移动电话号码和电话类型进行了分析，发现客户的性别、年龄、籍贯以及职业这四个人口统计变量在流失类别上有显著的差异。王维佳等人[27]利用决策树和RBF预测工具，以分类和预测概率作为目标，对客户流失进行了研究，并分析了影响客户流失的重要属性。Rosset（2003）[28]等人利用Logistic回归建立了客户流失预测模型，发现模型的提升率得到了较大的改善。Yang yiming（2004）等人[29]利用序列聚类方法，对客户流失进行了非监督分类。尹婷等人[30]提出了建立贝叶斯决策树算法的预测模型，将贝叶斯分类的先验信息和决策树分类的信息熵增益相结合，应用到电信行业客户流失预测中，得到了比单一采用决策树算法建立的流失预测模型具有更高的覆盖率与命中率。朱志勇等人[31]使用贝叶斯网络方法对客户进行流失分析，表明了基于Blanket模型的贝叶斯网络方法具有较高的预测精度。罗彬等人[32]在模型集成技术和代价敏感学习理论的基础之上，提出了基于蚁群算法的代价敏感线性集成多分类器的电信客户流失预警模型，其实验结果证明了该集成方法的可行性和有效性。Piew data[33]等人率先将人工智能技术应用于用户流失预测研究，他们的研究说明，人工神经网络算法能够有效、准确的预测用户的流失。hyunseok Hwang[34]分别使用了人工神经网络、决策树C4.5算法和Logistic回归算法对用户的利润贡献进行了分析预测，比较发现人工神经网络在预测精度上具有一定的优势。jorge等人[35]引入用户保留成本的概念，通过对神经网络、遗传算法和模糊神经系统的实际测试，发现神经网络方法较优。赵红星等人[36]提出了以粗糙集（RS）\_RBF神经网络作为电信业客户流失的预测模型，并使用实际数据证明了该模型的有效性。王学文等人[37]通过对电信客户流失的成因进行分析，采用改进的 BP 神经网络算法建立了电信业客户流失预测的神经网络模型，并使用该模型来预测未来客户的流失情况。田玲等人[38]提出了一种基于神经网络的客户流失预测模型。根据行业专家的经验选取分析变量，并通过神经网络计算所选分析变量的权值，建立了客户流失预测模型并对客户流失的趋势进行预测。蒋国瑞等人[39]将代价敏感学习应用于 Veropoulos 提出的采用不同惩罚系数的支持向量机，建立了客户流失预测模型，并采用了实际的电信客户流失数据进行验证，测试数据表明和传统的支持向量机、C4.5、ANN算法相比，该方法在准确率、命中率和覆盖率上均有提升。方磊和马溪骏[40]也在采用不同惩罚系数的支持向量机算法的基础上，使用样本信息熵来确定不同的惩罚因子，从而使模型更加倾向于提高少数类的识别准确度。邱一卉和林成德[41]提出了基于随机森林和转导推理的特征提取方法对数据集进行降维，并采用单类支持向量机算法建立了流失用户的预测模型。

在电信用户流失预警研究领域中，基于多分类器组合的相关研究还比较少，在实际中也没有得到应用，王纯麟等人[42]针对单分类器模型的不足之处，建立了一种基于AdaBoost组合分类器的电信客户流失预测模型，测试数据表明该模型取得了较好的结果。征荆等人[43]提出了一种基于误判代价和时间代价最小准则的分类器动态集成方法，动态地为每一测试样本选择最合适的一组分类器进行集成。陈晔和覃晓群[44]采用组合预测的理论知识，提出了一种基于组合预测理论的预测算法，并对电信客户流失预测采用等权平均法、方差-协方差法、最优加权法、递归等权法和递归方差倒数法分别建立组合预测模型，对客户进行了更准确的预测。孟飞翔[45]选择决策树和神经网络作为基本分类算法，通过Bagging或者Boosting算法进行集成，结果表明C4.5加上AdaBoost建立的客户流失预测模型具有最好的分类性能。

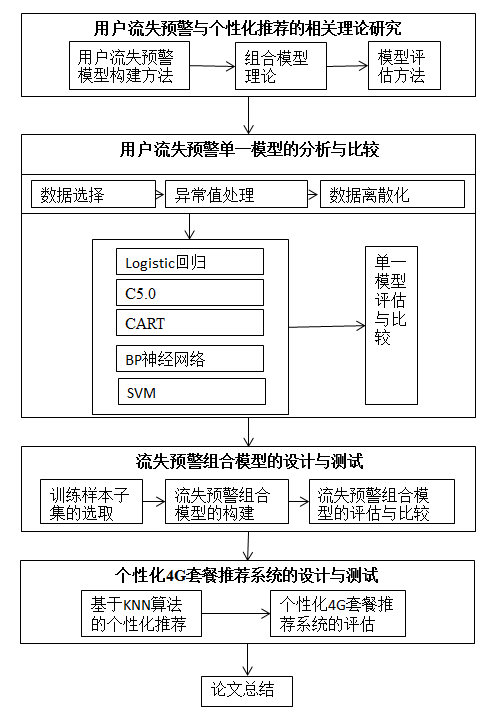
尽管研究用户流失预警问题的研究人员和企业工作者均做了大量的研究，但是依然存在一些问题。比如实际用户的数据不对称，评价标准单一，用户历史数据不完整等等，使得所建立的模型很难在实际中得到应用，特别是建模数据的选取通常采用的是抽取一些质量很高的用户数据来建立流失预警模型，而这些数据在实际中对大部分用户数据而言，都是很难获取或者是数据存在虚假问题。比如几乎所有流失预警模型都选取的性别和年龄字段，而这两个字段在实际中对大部分存量用户而言都缺失并且即使获取到该字段信息的用户的数据大部分都是虚假的。甚至还有些流失预警模型将用户的职业和收入也纳入模型考虑，这两种类型的字段可能是手工获取的，在运营商系统平台上面根本就不存在，以上种种问题限制了现有的流失预警模型在我国电信行业的应用。

个性化推荐在当前技术已经比较成熟，但是在我国通信行业并没有得到大规模的应用。一是由于运营商通常是从短期盈利角度考虑，忽略了用户感知；二是由于通信行业用户的敏感程度要低于电子商务行业，所以当运营商采用新的技术手段来改善用户体验时比较谨慎。三是由于通信行业的用户处于被锁定的状态，更换号码的代价非常巨大。这样导致的后果就是尽管我国4G网络已经发展了2年，但是对用户进行4G套餐的推荐依然处于一种被忽视的状态。

## 1.3 技术路线与内容结构

### 1.3.1 技术路线

本文研究的技术路线如图1.2所示，流失用户预测一般被看做是一种对海量用户数据进行挖掘，从而获取相关的模式或者规则的过程，利用该模式或者规则可以判断用户在未来一段时间内是流失还是正常，如果该用户未来一段时间内会流失，则对其实施相应的维系挽留策略。由于进行个性化的4G套餐推荐的前提和核心是找到即将流失的用户。因此，流失用户的识别是本文研究的重点。本文所使用的数据来源于联通A省分公司BSS平台上的数据，由于BSS平台上的用户数据非常复杂，本文只选取可以从BSS平台上容易获取、真实并且不需要进行另外的接口开发的用户数据。因此，本文的研究具有普遍的意义。



##### 图1.2 本文技术路线

### 1.3.2 内容结构

本文共有六章，每章内容如下：

第1章，绪论。本章主要介绍了本文的研究背景和本文的研究所具有的现实意义，然后对国内外学术界和企业界对用户流失预警的研究进行了描述，然后说明了本书的技术路线和内容结构。

第2章，流失预警与个性化推荐的相关理论研究。本章首先对数据挖掘中分类算法在用户流失预警中的应用进行了说明，然后对组合模型的优势及可行性、组合模型中基分类器的选取策略进行了研究，最后对模型的评估方法进行了阐述。

第3章，用户流失预警单一模型的分析与比较。本章主要对当前在用户流失预警中所使用的主流的分类算法在用户流失预警中的应用进行了详细的描述，并将每种分类算法应用于当前的数据建立模型，然后对用户流失预警模型的性能进行了比较分析。

第4章，流失预警组合模型的构建与测试。本章首先采用两步聚类算法对用户数据进行聚类，形成有较大差异的训练样本子集，然后将前一章所分析的每一种分类算法应用于训练样本子集，形成基分类器，再对基分类器采用多数投票法构建用户流失预警组合模型。并且在本章中也将本文所构建的用户流失预警组合模型从模型的准确率、精确率、召回率和稳定度等方面与上一章采用单一分类算法所建立的用户流失预警组合模型进行了对比分析。

第5章 个性化4G套餐推荐系统的构建。本章以A省分公司在2015年12月从2G&3G 转4G的用户数据为基础，采用KNN算法建立个性化4G套餐推荐系统，从而形成了流失用户的维系挽留策略。

第6章，结论。本章对本文的研究内容进行了总结，指出了本文的研究成果和研究意义。

# 第二章 用户流失预警与个性化推荐的相关理论基础

电信行业用户流失是指用户因为某些原因停止使用电信运营商提供的服务的行为。用户停止使用电信运营商的服务的原因有很多, 但归纳起来主要有以下几类：一是自然流失，是指运营商不能提供用户所期望的服务或者是由于某些客观的因素（如工作地点的变动）而选择离网所导致的用户流失；二是恶意流失，是指用户因为个人私欲因素（如恶意欠费）选择离网而导致的用户流失；三是竞争流失，是指竞争对手对该用户提供更合适的服务（如更优惠的套餐价格）而促使用户选择离网所导致的用户流失。

显然，对即将流失的用户在离网之前进行预警将会对电信企业而具有非常重要的价值。若能预测用户将自然流失，则运营商可通过完善产品和服务以满足用户新的需求, 从而挽留住该类用户户。若能预测用户户将恶意流失, 则运营商可及时终止服务或采取法律手段追回欠费以减少坏账损失。若能预测用户将竞争流失, 则运营商可及时调整资费政策以防止用户离网。

## 2.1 用户流失预警建模所采用的技术

电信行业的流失预警问题可以看成是数据挖掘中的分类问题，即判断用户是在下月是处于正常服务状态或者是流失状态。因此，该问题的本质是一种二分类问题。早期的流失预警模型以统计学模型为主，最主要的两种方法就是logistic回归算法和决策树算法，这两种方法都得到了广泛的应用并在实际中取得了良好的效果。从20世纪90年代开始，随着计算机技术和人工智能的迅速发展，支持向量机和人工神经网络也开始应用于用户流失预警的研究，并且方法日趋成熟。

（1） logistic回归

logistic回归[46]模型是一种广义线性回归（generalized linear model）模型，因此与多重线性回归分析有很多的相同之处。它们的模型形式上基本相同，都具有 w·x + b的形式，其中w和b是待求参数，其区别在于他们的因变量不同，多重线性回归直接将w·x +b作为因变量，即y = w·x +b，而logistic回归则通过函数L将w·x +b对应一个隐状态P，P =L(w·x +b),然后根据P与1-P的大小决定因变量的值。Logistic回归模型的表现形式如式（2.1）。

 （2.1）

在电信行业用户流失预警模型中，表示为用户流失的概率，当大于0时，大于0.5，说明用户流失的概率大于0.5，即预测此用户将会流失。

（2）决策树

决策树算法是一种有指导的数据挖掘方法，是使用最广泛的分类预测方法之一，常用的决策树算法包括ID3[47]算法、C4.5算法[48]、CART算法[49]，C5.0算法则优化了C4.5算法的执行效率和内存使用，其核心思想和C4.5算法一致。决策树呈树形结构，一颗决策树通常由一个根节点、若干个分支节点和若干个叶节点组成，叶节点对应于分类决策结果。其他的每个节点对应于一个属性上的测试，每个节点中的样本集合根据属性的测试结果划分到相应的子节点中，决策树的建立过程就是决策树对训练样本集不断分组的过程，也就是各个分支依次形成的过程，决策树的核心问题就是如果从众多的输入变量中选择一个最佳的分组变量，依据算法的不同，选择最佳分组变量的方式也不相同，ID3算法和C4.5算法分别采用信息增益和信息增益率选择分组变量，CART算法则采用基尼系数来选择分组变量。

（3）支持向量机

支持向量机算法[50]是建立在统计学习理论的VC维理论和结构风险最小化理论基础上的，根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折中，以求获得最好的推广能力。

（4）人工神经网络

神经网络[51](Artificial Neural Network,ANN)是由具有适应性的简单单元组成的广泛并联的网络，它是一种模拟人脑思维的抽象计算模型。神经网络中最基本的成分是神经元模型，也称为处理单元（processing element）或者节点，完整的节点通常是由加法器和激活函数组成。节点之间的连接称为边，边反映了各节点之间的关联性，而关联性的强弱体现在边的权重上，神经网络的学习就是根据学习经验不断调整权值的过程。神经网络的种类非常多，按照拓扑结构 ，神经网络可以分为两层神经网络、三层神经网络以及多层神经网络；按照神经网络的层间连接方式，神经网络可分为前馈式神经网络和反馈式神经网络。在实际应用中使用神经网络模型，大多采用的是BP神经网络。

Logistic回归、决策树、神经网络和支持向量机是当前在电信行业用户流失预警最常用的算法，本文对流失预警组合算法的研究及对比均以此四种算法为基础。同时，由于决策树模型有很多种形式，但是最常用的是C4.5和CART算法，本文对决策树的应用研究也将以此两种算法来进行。

## 2.2 组合模型概述

### 2.2.1 组合分类器的优势

确定一组合理的候选分类模型，采用带有类别标签的样本集对候选的分类器进行训练，然后用独立的的测试数据集集估计候选分类器的性能，最后选择泛化性能最好的分类器，这是实际中分类器设计最常用方法。用这种方法可以得到应用在整个特征空间上的单个最佳分类器。但是模型的好坏的评价通常有很多指标，比如准确率、精确率、召回率和稳定性等，在实际中可能会经常出现这样一种情况，在一组分类器中，没有哪一种分类器是最优的，也就是不能找到一个在所以评价指标上均优于其他模型。然而，某个分类器的错分样本集与另一个分类器的错分样本集一般不会完全一致，因此不同的分类器可以给出互补的信息，于是将这样的分类器组合起来是有益的。将分类器组合起来并不是一种最新思想，但是这种做法在近年来获得了越来越多的关注。

多分类器组合是为了提高分类性能尤其是分类正确率而提出的一种技术，多分类器组合是一个包含了多个不同的分类器的组合，其中每个分类器称为基分类器。每个基分类器都参与对测试数据的分类并给出一个分类结果，然后按照分类器组合方法组合这些分类结果并给出测试数据集中每一实例的最终分类结果。实验数据表明，多分类器组合与单分类器相比，能显著提升分类算法的分类精度。多分类器组合比其中任何一个基分类器分类精度都高的一个充要条件是基分类器是准确且多样的，基分类器的准确是指该分类器在测试数据集中的分类错误的概率应该比随机猜想要小，在实际应用中，基分类器的准确率一般都会满足此种要求。基分类器的多样性是指在测试数据集中，各基分类器的分类结果应该尽可能是不同且独立的，尤其是对分类错误的实例而言。采用多分类器组合之所以可以构建出比单分类器具有更好的准确率的模型的原因有以下三点：

（1）从统计角度分析，一个学习算法实际上可以看成是在一个假设空间上寻找一个最好的并且跟实际假设最接近的一个假设。但是在通常情况下，训练样本相对于整个假设空间是不充分的，所以通常在当前数据集上找到的比较准确的假设并不一定接近真正的假设，如果组合多个这样的假设，学习算法就可以平均化它们的结果，从而减少错误分类的概率，最终得到的假设可以更加接近真实的假设。

（2）从计算的角度分析，许多实际中的学习算法都是通过在局部假设空间中进行搜索来工作的，这样学习算法就可能陷入局部最优而非全局最优的缺陷中。例如，神经网络算法一般采用了梯度下降方式来减少其在训练数据集上的分类错误率；决策树算法采用贪心策略来生成决策规则，这样即使训练数据集样本量非常充分，学习算法也只是在有限的局部假设空间内搜索，从而很难得到最好的假设。通过组合多个学习算法从不同的点开始局部搜索，这样就有可能搜索到比任何一个学习算法寻找到的局部最优假设跟真实假设更接近的假设。

（3）从表示的角度分析，在大多数的机器学习任务中，真实假设不能被搜索空间中的任何一个假设所表示，而通过组合多个从假设空间中得到的假设，可能会扩展假设空间，即增强泛化能力。

### 2.2.2 基分类器的选择及组合

多分类器组合实际上也是一个从实例到类别的映射，假设训练数据集为，其中表示第k个训练实例，表示第k个训练实例所属的类别，多分类器组合为，其中为第k个基分类器，每一个及分类器都是一个从实例到类别的映射：，多分类器组合即可表示为：，也就是使用某种组合方法组合多个基分类器的分类结果从而形成一个最终的分类预测。

由上述的概念可知，多分类器组合主要包括两个方面：一方面是基分类器的生成阶段，即如何生成多个基分类器；另一个方面是基分类器的组合阶段，即采用何种方式组合基分类器的分类结果来对实现对实例的分类。

在实际中，通过对训练数据集的选择、学习算法的选择或者算法参数的设置，可以很轻松的构建成百上千个基分类器，但是这种方法会产生很多的负面影响：一方面，使用更多的分类器将导致更大的计算和存储开销；另一方面，当基分类器数目增加之后，分类器之间的差异性可能会很小。因此，并不是组合的基分类器的数目越多越好。

差异性被认为是制约多分类器组合性能差异的一个关键因素，所谓分类器的差异性是指对测试实例进行分类时，基分类器的分类结果是独立且不同的，这样即使有某个基分类器对测试实例做出错误的分类预测，也会有其它的基分类器来纠正这个错误，这样就有可能提高分类器的分类性能。衡量差异性的方法被称为差异性度量方法 （Diversity Measure）。常用的分类器差异性度量方法有Q 统计量方法、不一致度量方法、协方差度量方法等。

两个基分类器 Di和 Dj在训练样本集中的分类结果如表2.1所示：

#### 表2.1 两个基分类器分类结果

|  |  |  |
| --- | --- | --- |
|  | 被Dj正确分类 | 被Dj错误分类 |
| 被Di正确分类 | N11 | N10 |
| 被Di正确分类 | N01 | N00 |

其中N11与N00代表两个基分类器均预测正确与均预测错误，即两基分类器均作出正确预测或错误预测的训练样本占总训练样本的比例。N10为Di预测正确而在Dj中预测错误，N01为Di预测错误而在Dj中预测正确。

（1）Q统计量

Q统计方法对两个基分类器Di和 Dj之间的差异性，Q统计量定义如式（2.2）：

 （2.2）

Q统计量用来测量两个分类器之间的差异度，的大小与两个基分类器分类结果相同个数的大小成正比，并且取值范围为[-1,1]，的绝对值越大，两基分类器的差异性越小。对于独立的基分类器来说，Q的期望值应该为0，整个分类器组合的Q统计量是所有两两基分类器的Q统计量的均值，表达式如式（2.3）：

 （2.3）

（2）不一致度量

不一致度量方法对两个基分类器Di和 Dj之间的差异性定义如下：

 （2.4）

整个分类器组合的差异性度量是所有两两基分类器的的均值：

 （2.5）

不一致度量值变化的范围从0到1之间。当两个基分类器同时将每一个目标分类正确或错误时，度量值会是0，当两个分类器预测不同，且有一个预测是正确时，度量值会是1。越大，分类器间差异性越大。

（3）协方差度量

于两个基分类器来说，协方差P的度量公式如下：

 （2.6）

协方差P与Q统计量比较类似，整个分类器组合的协方差是所有两两基分类器的协方差P的均值，的取值如下：

 （2.7）

协方差P越小，基分类器之间的的差异性越大。

在确定了多分类器组合的基分类器后，接下来需要考虑的问题就是如何将这些基分类器对测试实例的分类结果进行组合，从而构成对测试实例的最终分类。对一个测试实例x，多个基分类器对其的分类结果如表2.2所示：

#### 表2.2 基分类器组合分类结果表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Y1 | Y2 | …… | Yn |
| D1 | P11 | P12 | …… | P1n |
| D2 | P21 | P22 | …… | P21 |
| …… |  |  |  |  |
| Dm | Pm1 | Pm2 | …… | Pmn |

表2.2中，Yn代表第n个类别，Dm代表第m个基分类器，Pij表示的是基分类器Di将x判断为类别yj的概率，且，如果基分类器Di将x判断为yj，并且x的实际类别也为yj，则Pij=1，否则为0。

常见的基分类器组合方法有：多数投票法、性能加权方法、分布求和法、贝叶斯联合法、熵加权法等，电信用户流失预警的实质是一个二分类问题，本文采用多数投票法作为基分类器的组合方法。多数投票法是最常用的一种组合方法，它适用于分类结果为仅给出类别标签的情形，多数投票法统计每个类值被预测的次数，被预测次数最多的类值被认为是测试实例的类值，假设Y1代表用户下月流失，Y2代表用户下月正常在网，如果，则判断为该用户下月将流失，否则判断用户下月依然正常在网。为了解决时无法判断用户属于哪一个类别的问题，m一般取奇数，也就是说多分类器组合一般选择奇数个基分类器。

## 2.3 基于4G套餐的个性化推荐方法

个性化推荐是指建立在海量数据挖掘基础之上根据用户的消费行为特征或者是兴趣特点，挖掘出用户的潜在偏好，从而向用户推荐其感兴趣的商品或者是信息。体现在通信行业的4G套餐中就是根据用户的属性和消费行为（用户的语音使用量、短信使用量、流量使用量和出账金额等），向用户推荐其最感兴趣的4G套餐。显而易见，对运营商而言，向用户推荐4G套餐最直观的推荐技术就是协同过滤算法，协同过滤算法可以分为两类：一类是基于用户的协同过滤，此种推荐方法是通过不同用户对物品的评分来比较用户之间的相似性，然后基于用户之间的相似性来做出推荐决策；另一类是基于物品的协同过滤，此种推荐方法是通过不同用户对物品的评分来评测物品之间的相似性，然后基于物品之间的相似性做出推荐决策。

假设用户的通信消费数据用向量表示，表示用户的消费特征上的消费值（比如上网流量，主叫时长等），那么用户和用户的相似性即可由向量和向量之间的相似性表示。如果将物品看成是消费特征，协同过滤算法中对物品的评分在4G套餐推荐中自然转化成了用户在消费特征上的消费值，从而，基于4G套餐的个性化推荐也就转化成了基于用户的协同过滤算法解决的问题。

## 2.4 模型评估

模型评估是数据挖掘任务中非常重要的一环。不同的数据挖掘任务有着不同的评价指标，同时同一种数据挖掘任务也有着不同的评价指标，每个指标的着重点不一样。如分类（classification）、回归（regression）、排序（ranking）、聚类（clustering）、热门主题模型（topic modeling）、推荐（recommendation）等。并且很多指标可以对多种不同的数据挖掘模型进行评价，如精确率－召回率（precision-recall），既可以用在分类中，也可以用在推荐和排序中。

（1）准确率（Accuracy）

准确率是数据挖掘中对分类模型评估最经常使用的一种评估方式，它是指在分类中，使用测试集对模型进行分类预测时，分类正确的实例个数占总实例个数的比例：

 （2.8）

式中表示模型判断正确的实例的个数，表示总的实例的个数。

准确率是从所有测试样本的角度来评估模型，没有对不同类别进行区分，即其平等对待每个类别，因此这种评估方式在很多应用场景中存在问题，比如在本文的用户流失预警模型中，对流失用户和正常用户的分类的错误的代价是不同的，而准确率却不能反映这一问题。还有一个原因就是，样本中不同类别的样本可能分布不平衡，也就是有的类别下的样本过多，有的类别下的样本个数过少，当所属不同类别的实例个数相差较大时，将会导致样本占大部分的类别主导准确率的计算。

（2）平均准确率(Average Per-class Accuracy)

平均准确率是为了在每个类别样本的个数不一致的情况下对模型进行评估，在准确率的基础上，计算每个类别下的准确率，然后再计算它们的平均值。如果每个类别下类别的样本个数不一致，则平均准确率不等于准确率；如果每个类别下的样本个数一致，则平均准确率与准确率相等。平均准确率的定义如式2.9：

 （2.9）

式（2.9）中n表示样本中的类别个数，表示测试集中模型判断为类别i的个数，表示模型判断为类别i并且该样本实际也为类别i的样本数。

平均准确率的缺点是如果存在某个类别，类别的样本个数很小，那么使用测试集进行测试时，可能造成该类别准确率的方差过大，意味着该类别的准确率可靠性不强。

（3）精确率-召回率(Precision-Recall)

精确率-召回率准确来说是两个评估指标，但是它们一般都是同时使用。精确率是指分类器分类正确的正样本的个数占该分类器所有分类为正样本个数的比例。召回率是指分类器分类正确的正样本个数占所有的正样本个数的比例。

精确率的定义如下：

 （2.10）

召回率的定义如下：

 （2.11）

在精确率和召回率的定义中，表示分类器分类正确的正样本的个数，表示所有分类为正样本的个数，表示所有测试集中正样本的个数。

由精确率和召回率可以求出F1-score，F1-score为精确率与召回率的调和平均值，它的值更接近于Precision与Recall中较小的值。

 （2.12）

（4）混淆矩阵(Confusion Matrix)

混淆矩阵是对分类的结果进行详细描述的一个表，对于二分类则是一个2\*2的矩阵，对于n分类则是一个n\*n的矩阵。对于二分类而言，第一行是真实类别为正的实例个数，第二行则是真实类别为负的实例个数，第一列是预测值为正的实例个数，第二列则是预测值为负的实例个数。以二分类的混淆矩阵为例，混淆矩阵的结构如表2.2所示：

#### 表2.3 混淆矩阵的结构

|  |  |  |
| --- | --- | --- |
|  | 预测为正 | 预测为负 |
| 实际为正 | TP | FN |
| 实际为负 | FP | TN |

表2.2中的混淆矩阵可以将模型的测试结果分为四类：

真正(True Positive, TP)：被模型分类正确的正样本。

假负(False Negative, FN)：被模型分类错误的正样本。

假正(False Positive, FP)：被模型分类的负样本。

真负(True Negative, TN)：被模型分类正确的负样本。

通过混淆矩阵可以计算模型评估中常用的评估指标，如准确率、平均准确率、精确率、召回率和F1-score，各评估指标的计算方法如下：

准确率(Accuracy)：

 （2.13）

平均准确率(Average Per-class Accuracy)：

 （2.14）

精确率(Precision)：

 （2.15）

召回率(Recall)：

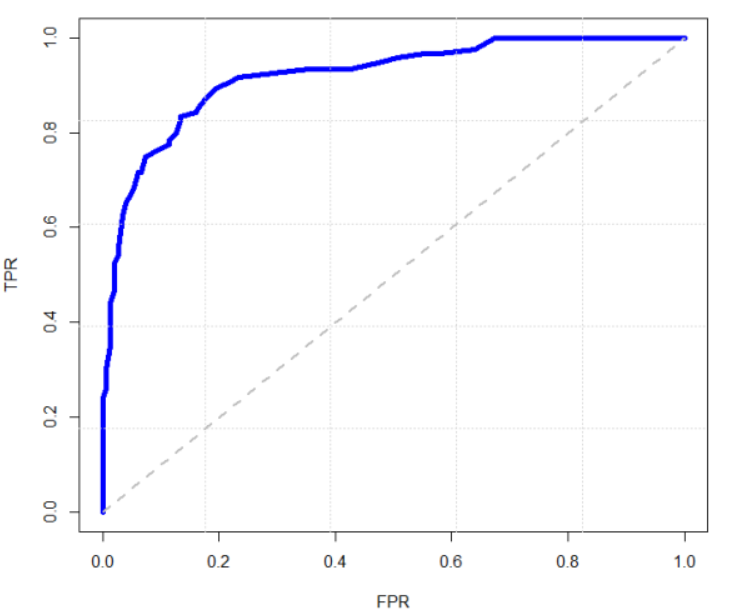
 （2.16）

F1-score：

 （2.14）

（5）ROC曲线

ROC曲线全称是接受者操作特征曲线。ROC曲线描述分类器的True Positive Rate（TPR，分类器分类正确的正样本个数占总正样本个数的比例）与False Positive Rate（FPR，分类器分类错误的负样本个数占总负样本个数的比例）之间的变化关系。ROC曲线图2.1。



##### 图2.1 ROC曲线

ROC曲线描述了FPR不断变化时的TPR的值，即ROC曲线是FPR与TPR之间的关系曲线。在ROC曲线中，点(0,0)表示TPR＝0并且FPR＝0，即分类器将每个实例都预测为负类。点(1,1)表示TPR＝1并且FPR＝1，即分类器将每个实例都预测为正类。点(0,1)表示TPR＝1并且FPR=0，即分类器将每个正类实例都预测为正类，将每个负类实例都预测为负类。

显而易见，最好的分类器便是FPR＝0%并且TPR＝100%，但是一般在实践中一个分类器很难会有这么好的效果，即一般TPR不等于1，FPR也不等于0。ROC凸壳可以作为一个识别潜在的最优分类器的鲁棒性的方法，若一个分类器的ROC曲线被另一个分类器的ROC曲线完全包住，则认为该分类器的效果比另一分类器差，若两个分类器的ROC曲线相交叉，则可通过ROC曲线下方的面积进行比较，面积较大的分类器的分类效果较优。

# 第三章 用户流失预警单一模型的分析与比较

电信行业用户流失预警问题实质是一种二分类问题，在实际中几乎所有的能解决分类问题的方法均可用于对用户流失预警建模。现有的用户流失预警模型大致可以分为统计模型和非统计模型，统计模型主要包括有多元线性回归、logistic回归，判别分析、决策树和KNN算法等，它们在流失预警中得到广泛的应用并取得了很好的效果，并且在当前依然处于主流地位，主要是因为此类模型具有复杂度低、稳定性好并且可解释性强的优点，缺点在于此类方法通常都对数据具有较严格的要求，例如logistic回归要求自变量之间没有很强的相关性。在流失预警中常用的非统计模型主要是支持向量机和人工神经网络，此类方法通常具有分类精度高并且对数据没有严格的要求的优点，缺点是此类方法在建模时的复杂度比较高，而且采用此类方法建立的模型通常缺乏解释性。本章拟采用联通A省份公司的用户数据对在电信行业用户流失预警中最常用的四种分类算法进行了实证分析，并且比较了各种分类算法在准确率、精确率、召回率和稳定性方面的差异，从而使本章成为下章构建流失预警组合模型的基础。

## 3.1 用户数据的选择

本文研究所选取得数据来自于联通A省分公司2015年6月至11月的有正常消费（排除无主叫、无被叫、无短信且无上网流量）的用户数据，目标变量为判断用户在2016年1月份是否流失。本文在变量的选取中排除了一些在其他流失预警中经常用到的数据，例如：用户性别、用户年龄等，主要是因为在和省分公司的交流中了解到这些数据存在大量的缺失数据，并且最重要的是非缺失数据中也存在很大比率的虚假数据；同时本文也没有采用用户日时间段的数据，因为此部分数据的提取需要对运营商现有的系统进行改造，这也是流失预警模型在理论界得到了比较充分的研究，但是在实际中很难得到应用的最主要的原因。本文初始所提取的字段信息如表3.1所示：

#### 表3.1 初始字段信息提取表

|  |  |  |
| --- | --- | --- |
| 变量名称 | 变量类型 | 变量意义 |
| USER\_NO | 数值(N) | 用户编码 |
| SERVICE\_KIND | 字符串 | 业务类型 |
| AREA\_DESC | 字符串 | 地市名称 |

#### 续表3.1 初始字段信息提取表

|  |  |  |
| --- | --- | --- |
| 变量名称 | 变量类型 | 变量意义 |
| GROUP\_FLAG | 字符串 | 客户类型 |
| VIP\_TYPE | 字符串 | VIP类型 |
| CHANNEL\_TYPE | 字符串 | 渠道类型 |
| RENT\_FEE | 数值(N) | 基本月租 |
| INNET\_MONTH | 数值(N) | 在网时长 |
| SALE\_KIND | 字符串 | 合约类型 |
| IF\_WO\_USER | 字符串 | 是否沃家庭 |
| SCORE\_VALUE | 数值(N) | 当前积分 |
| TOTAL\_FEE\_AVG | 数值(N) | 平均出账收入 |
| QF\_FEE\_AVG | 数值(N) | 平均累计欠费 |
| LEFT\_FEE\_AVG | 数值(N) | 平均预存款余额 |
| CDR\_02\_NUM\_AVG | 数值(N) | 平均异网通话次数 |
| CDR\_01\_NUM\_AVG | 数值(N) | 平均同网通话次数 |
| JF\_01\_TIMES\_AVG | 数值(N) | 平均本地主叫计费时长 |
| JF\_02\_TIMES\_AVG | 数值(N) | 平均本地被叫计费时长 |
| GPRS\_AVG | 数值(N) | 平均上网流量 |
| SMS\_NUM\_AVG | 数值(N) | 平均短信数量 |
| BIN\_STOP\_AVG | 数值(N) | 平均半停次数 |
| ALL\_STOP\_AVG | 数值(N) | 平均双停次数 |
| TOTAL\_FEE\_DEV | 数值(N) | 出账收入离均差 |
| QF\_FEE\_DEV | 数值(N) | 累计欠费离均差 |
| LEFT\_FEE\_DEV | 数值(N) | 预存款余额离均差 |
| CDR\_02\_NUM\_DEV | 数值(N) | 异网通话次数离均差 |
| CDR\_01\_NUM\_DEV | 数值(N) | 同网通话次数离均差 |
| JF\_01\_TIMES\_DEV | 数值(N) | 本地主叫计费时长离均差 |
| JF\_02\_TIMES\_DEV | 数值(N) | 本地被叫计费时长离均差 |

#### 续表3.1 初始字段信息提取表

|  |  |  |
| --- | --- | --- |
| 变量名称 | 变量类型 | 变量意义 |
| GPRS\_DEV | 数值(N) | 上网流量离均差 |
| BIN\_STOP\_DEV | 数值(N) | 半停次数离均差 |
| ALL\_STOP\_DEV | 数值(N) | 双停次数离均差 |
| LEAVE\_NET | 数值(N) | 是否流失 |
| TOTAL\_FEE\_STD | 数值(N) | 出账收入标准差 |
| QF\_FEE\_STD | 数值(N) | 累计欠费标准差 |
| LEFT\_FEE\_STD | 数值(N) | 预存款余额标准差 |
| CDR\_02\_NUM\_STD | 数值(N) | 异网通话次数标准差 |
| CDR\_01\_NUM\_STD | 数值(N) | 同网通话次数标准差 |
| JF\_01\_TIMES\_STD | 数值(N) | 本地主叫计费时长标准差 |
| JF\_02\_TIMES\_STD | 数值(N) | 本地被叫计费时长标准差 |
| GPRS\_STD | 数值(N) | 上网流量标准差 |
| SMS\_NUM\_STD | 数值(N) | 短信数量标准差标准差 |
| BIN\_STOP\_STD | 数值(N) | 半停次数标准差 |
| ALL\_STOP\_STD | 数值(N) | 双停次数标准差 |
| ALTER\_SERVICE\_KIND | 数值(N) | 业务类型是否变化 |
| ALTER\_F\_PROD\_ID | 数值(N) | 套餐编码是否变化 |
| ALTER\_GROUP\_FLAG | 数值(N) | 客户类型是否变化 |
| ALTER\_VIP\_TYPE | 数值(N) | VIP类型是否变化 |
| ALTER\_RENT\_FEE | 数值(N) | 基本月租是否变化 |
| ALTER\_SALE\_KIND | 数值(N) | 合约类型是否变化 |
| ALTER\_IF\_WO\_USER | 数值(N) | 是否沃家庭是否变化 |

本文选取如表3.1所示的用户数据，其中USER\_NO变量用来标示唯一用户，平均值数据代表此类变量在2015年6月份到2015年11月份6个月内的平均值，代表该变量的集中趋势；标准差数据代表此类变量在2015年6月份到2015年11月份6个月内的标准差，代表该变量的离散趋势；离均差数据代表此类变量在2015年11月份的数据与此类数据平均值的偏离程度，代表此类变量在11月份的变化趋势；同时也将类别型数据和用户的基本月租在11月份和10月份的进行对比，看其是否发生变化，从而也将其以变量的形式纳入模型考虑。

本文初步选取以上51个变量纳入考虑，除了用户的唯一标示USER\_NO和目标变量LEAVE\_NET外，在实际中，对用户流失预警模型而言，有一些变量和用户流失是无关的，需要进行变量选择。所谓变量选择，就是从众多的输入变量中，找出对输出变量分类预测有意义的重要变量。因此，如果要建立用户流失预警模型，那么在建模前进行变量选择是必要的，如果不经过变量选择，所有输入变量全部参与建模，不仅会影响模型的计算效率，更重要的是，由于输入变量之间可能存在相关性等，会使所得模型无法用于预测。

变量选择的主要目的是寻找对输出变量预测有重要作用的变量，变量的重要性可以从两个方面联合考察：第一，从变量本身考察；第二，从输入变量与输出变量的相关性角度加以考察。

从变量本身的角度考虑，重要的变量应该是携带信息较多，也就是变量值差异较大的变量。统计上，测度数值型变量取值离散性的指标是标准差或者是变异系数。标准差越大，说明变量取值的离散程度越大，反之，越小。变异系数在消除数量级影响的情况下，便于对多个变量的离散程度进行对比，变异系数的计算公式为：

变异系数 CV =( 标准偏差 SD / 平均值Mean )× 100% （3.1）

对数值型变量重要性的判断标准通常有以下两种方式：

（1）如果数值型变量的变异系数小于某个阈值，则该变量被视为不重要的变量。

（2）如果某数值型变量的标准差小于某个阈值，则该变量被视为不重要的变量。

本文对数值型变量的重要性的判断标准采用的是变异系数来确定，当变异系数小于0.1时，该变量为不重要的变量，当变异系数大于等于0.1时，该变量为重要的变量。

对分类变量而言，本文对变量的考察标准是：单个类别中记录的最大百分比为90%；作为记录百分比的最大类别数为95%。

从输入变量与目标变量之间的相关性角度考察，对于不同类型的变量，测度相关性的方法存在差异，在用户流失预警模型中，由于目标变量为分类变量，因此变量的选择可以以输入变量的类型分别加以考虑：

（1）分类变量。当输入变量为分类变量时，可以采用卡方检验来对输入变量进行检验，卡方检验属于非参数检验的范畴，在本文中，卡方检验采用似然比卡方（likelihood ratio），其数学定义为：

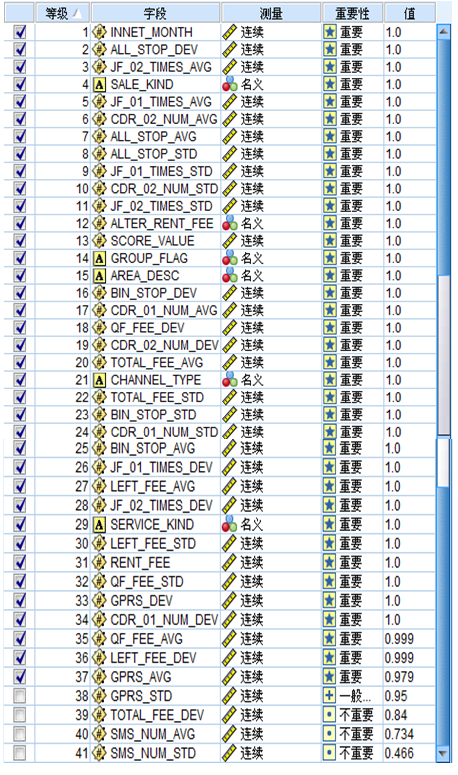
 （3.2）

式（3.2）中，为实际频数，为理论频数。本文计算似然比卡方的实际值和对应的概率P值，如果P值越小，输入变量与目标变量之间的总体线性相关的把握就越大，也就是说输入变量对目标变量的预测的重要意义。

（2）输入变量为数值型变量。此种情况下通常可以采用方差分析方法，如果目标变量不同类别水平下输入变量的均值存在显著性的差异，则输入变量和目标变量之间的相关性较强，反之则相反。本文采用计算方差分析中F检验统计量的观测值和对应的概率P值，P值越小，输入变量与目标变量相关的把握越大，输入变量对目标变量的预测有重要意义。

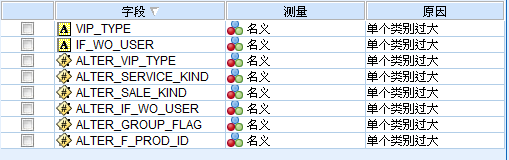
本文对输入变量的重要性判断的标准采用SPSS Modeler默认的系数，即判断输入变量很重要、中等重要和不重要的标准值为1-概率P值大于0.95为重要，在0.9~0.95之间为中等重要，小于0.9为不重要。

本文选取的用户数量为42161个用户，其中正常用户数量21393，流失用户数量20768。本文采用IBM SPSS Model 14.1进行变量的选择，输入变量与目标变量的相关性的检验结果如图3.1。



##### 图3.1 输入变量与目标变量相关性检验

从输入变量本身的角度考虑，分类变量频数检验结果如图3.2。



##### 图3.2 分类变量频数检验

综合以上分析，对构建用户流失预警模型而言，所有作为模型的输入变量总共49个（USER\_NO为用户的唯一标示，只作为用户标示，不参与建模。LEAVE\_NET为目标变量），最终纳入模型考虑的输入变量的个数为37个，不纳入模型12个。综合以上分析，建立用户流失预警模型所选择的变量如表3.2：

#### 表3.2 流失预警模型所选变量

|  |  |  |  |
| --- | --- | --- | --- |
| 变量名称 | 变量类型 | 变量意义 | 变量作用 |
| USER\_NO | 数值(N) | 用户编码 | 用户标示 |
| SERVICE\_KIND | 字符串 | 业务类型 | 输入变量 |
| AREA\_DESC | 字符串 | 地市名称 | 输入变量 |
| GROUP\_FLAG | 字符串 | 客户类型 | 输入变量 |
| CHANNEL\_TYPE | 字符串 | 渠道类型 | 输入变量 |
| RENT\_FEE | 数值(N) | 基本月租 | 输入变量 |
| INNET\_MONTH | 数值(N) | 在网时长 | 输入变量 |
| SALE\_KIND | 字符串 | 合约类型 | 输入变量 |
| SCORE\_VALUE | 数值(N) | 当前积分 | 输入变量 |
| TOTAL\_FEE\_AVG | 数值(N) | 平均出账收入 | 输入变量 |
| QF\_FEE\_AVG | 数值(N) | 平均累计欠费 | 输入变量 |
| LEFT\_FEE\_AVG | 数值(N) | 平均预存款余额 | 输入变量 |
| CDR\_02\_NUM\_AVG | 数值(N) | 平均异网通话次数 | 输入变量 |
| CDR\_01\_NUM\_AVG | 数值(N) | 平均同网通话次数 | 输入变量 |
| JF\_01\_TIMES\_AVG | 数值(N) | 平均本地主叫计费时长 | 输入变量 |
| JF\_02\_TIMES\_AVG | 数值(N) | 平均本地被叫计费时长 | 输入变量 |

#### 续表3.2 流失预警模型所选变量

|  |  |  |  |
| --- | --- | --- | --- |
| 变量名称 | 变量类型 | 变量意义 | 变量作用 |
| GPRS\_AVG | 数值(N) | 平均上网流量 | 输入变量 |
| BIN\_STOP\_AVG | 数值(N) | 平均半停次数 | 输入变量 |
| ALL\_STOP\_AVG | 数值(N) | 平均双停次数 | 输入变量 |
| QF\_FEE\_DEV | 数值(N) | 累计欠费离均差 | 输入变量 |
| LEFT\_FEE\_DEV | 数值(N) | 预存款余额离均差 | 输入变量 |
| CDR\_02\_NUM\_DEV | 数值(N) | 异网通话次数离均差 | 输入变量 |
| CDR\_01\_NUM\_DEV | 数值(N) | 同网通话次数离均差 | 输入变量 |
| JF\_01\_TIMES\_DEV | 数值(N) | 本地主叫计费时长离均差 | 输入变量 |
| JF\_02\_TIMES\_DEV | 数值(N) | 本地被叫计费时长离均差 | 输入变量 |
| GPRS\_DEV | 数值(N) | 上网流量离均差 | 输入变量 |
| BIN\_STOP\_DEV | 数值(N) | 半停次数离均差 | 输入变量 |
| ALL\_STOP\_DEV | 数值(N) | 双停次数离均差 | 输入变量 |
| LEAVE\_NET | 数值(N) | 是否流失 | 目标变量 |
| TOTAL\_FEE\_STD | 数值(N) | 出账收入标准差 | 输入变量 |
| QF\_FEE\_STD | 数值(N) | 累计欠费标准差 | 输入变量 |
| LEFT\_FEE\_STD | 数值(N) | 预存款余额标准差 | 输入变量 |
| CDR\_02\_NUM\_STD | 数值(N) | 异网通话次数标准差 | 输入变量 |
| CDR\_01\_NUM\_STD | 数值(N) | 同网通话次数标准差 | 输入变量 |
| JF\_01\_TIMES\_STD | 数值(N) | 本地主叫计费时长标准差 | 输入变量 |
| JF\_02\_TIMES\_STD | 数值(N) | 本地被叫计费时长标准差 | 输入变量 |
| BIN\_STOP\_STD | 数值(N) | 半停次数标准差 | 输入变量 |
| ALL\_STOP\_STD | 数值(N) | 双停次数标准差 | 输入变量 |
| ALTER\_RENT\_FEE | 数值(N) | 基本月租是否变化 | 输入变量 |

## 3.2 异常值的处理

由于偶然因素、人为因素或者数据变异等原因，常有一小部分样本数据与总体数据在特征、结构或相关性方面出现较大的差异，这部分数据被称之为异常值（outlier）。通信行业用户数据存在异常值是一个非常普遍的现象，由于本文的数据是从用户的日志信息中提取，因此一般情况下排除了人为因素的影响，但是由于某些特定的因素可能会导致用户的真实的消费数据出现异常，比如诈骗短信异常的短信数量，恶意偷跑流量软件导致的异常的上网流量等，还有养卡用户的异常消费数据，这些异常值都会引起模型分类精度和泛化能力的下降。对异常值的分析除了能提高流失预警模型的预测精度外，还可能挖掘出更多深层次、有价值的信息，比如疑似养卡用户的发现等。因此，异常值的处理在流失预警中的具有非常重要的作用。

异常值的检测方式可以从以下两个方面进行区分，第一，对用户数据在某一属性上的取值进行检测。当用户数据在该属性上的取值与绝大多数用户在该属性上的取值存在较大差异时，则可认为此部分用户数据为异常值；第二，对用户数据进行多属性上的检测。很多用户数据虽然从单一属性的取值上考虑不存在异常情况，但是部分属性间的取值结构却与整个样本数据集在此部分属性间的结构呈现较大的差异，此部分数据也可能为异常值。在通信行业中，用户消费行为的异常行为通常带来较差的用户体验（比如恶意APP造成的流量异常）或者是用户变为沉默用户，或者是用户居住或者工作地址的变动，这种情况通常和用户流失具有非常高的关联度。因此，这样的异常值也应该将其纳入模型加以考虑。

本文对单一属性异常值进行检测的是采用平均值加减三倍标准差的方法。假设属性变量X的数学期望和方差都存在（本文所选取的数值变量均满足此前提），对于任意总体X和任意的，有：

 （3.3）

当时，有

 （3.4）

即属性变量的取值与其数学期望的差的绝对值超过3倍标准差的概率约为11%，当样本量很大时，所有用户在X上的取值一般都服从正态分布，此时数据超过3倍标准差的概率则仅为0.27%，这种情形为小概率事件，即认为正常情况下X的取值不能超过这一范围，当X的取值不在此范围内时，则可认为该用户的数据异常。

对数值变量的异常值检验数据如表3.3。

#### 表3.3 数值变量的异常值检验

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段 | 最小值 | 最大值 | 平均值 | 标准差 | 异常值判断上限 | 异常值判断下限 |
| RENT\_FEE | 0 | 886 | 49.474 | 84.12 | 301.834 | -202.886 |
| INNET\_MONTH | 1 | 230 | 58.558 | 43.604 | 189.37 | -72.254 |
| SCORE\_VALUE | 0 | 27279 | 1017.28 | 1503.07 | 5526.488 | -3491.932 |
| TOTAL\_FEE\_AVG | -167.11 | 1327.08 | 76.005 | 94.594 | 359.787 | -207.777 |
| QF\_FEE\_AVG | -564.01 | 4666.67 | 3.893 | 74.477 | 227.324 | -219.538 |
| LEFT\_FEE\_AVG | 0 | 31377.61 | 227.971 | 490.558 | 1699.645 | -1243.703 |
| CDR\_02\_NUM\_AVG | 0 | 3888.83 | 114.487 | 156.09 | 582.757 | -353.783 |
| CDR\_01\_NUM\_AVG | 0 | 12479.33 | 46.476 | 122.329 | 413.463 | -320.511 |
| JF\_01\_TIMES\_AVG | 0 | 8615.67 | 127.364 | 190.542 | 698.99 | -444.262 |
| JF\_02\_TIMES\_AVG | 0 | 12637.5 | 153.752 | 225.28 | 829.592 | -522.088 |
| GPRS\_AVG | 0 | 360930.1 | 371.48 | 3626.438 | 11250.794 | -10507.834 |
| BIN\_STOP\_AVG | 0 | 78.5 | 0.332 | 0.792 | 2.708 | -2.044 |
| ALL\_STOP\_AVG | 0 | 9.67 | 0.13 | 0.251 | 0.883 | -0.623 |
| QF\_FEE\_DEV | -2020.2 | 1733.33 | 2.076 | 33.687 | 103.137 | -98.985 |
| LEFT\_FEE\_DEV | -4164.5 | 3974.91 | -29.784 | 166.763 | 470.505 | -530.073 |
| CDR\_02\_NUM\_DEV | -1456.2 | 1626.5 | -7.952 | 63.975 | 183.973 | -199.877 |
| CDR\_01\_NUM\_DEV | -11133 | 5806.5 | -4.745 | 93.099 | 274.552 | -284.042 |
| JF\_01\_TIMES\_DEV | -6079.7 | 2176.17 | -10.531 | 95.803 | 276.878 | -297.94 |
| JF\_02\_TIMES\_DEV | -11292 | 5995.5 | -11.153 | 131.271 | 382.66 | -404.966 |
| GPRS\_DEV | -68005 | 121799.8 | 37.744 | 1021.39 | 3101.914 | -3026.426 |
| BIN\_STOP\_DEV | -47.67 | 37 | 0.029 | 0.746 | 2.267 | -2.209 |
| ALL\_STOP\_DEV | -9 | 36.5 | 0.049 | 0.388 | 1.213 | -1.115 |
| TOTAL\_FEE\_STD | 0 | 1395.49 | 81.336 | 97.356 | 373.404 | -210.732 |
| QF\_FEE\_STD | 0 | 4788.88 | 5.34 | 79.323 | 243.309 | -232.629 |
| LEFT\_FEE\_STD | 0 | 31380.27 | 246.251 | 506.971 | 1767.164 | -1274.662 |
| CDR\_02\_NUM\_STD | 0 | 3893.6 | 120.991 | 161.196 | 604.579 | -362.597 |

#### 续表3.3 数值变量的异常值检验

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段 | 最小值 | 最大值 | 平均值 | 标准差 | 异常值判断上限 | 异常值判断下限 |
| CDR\_01\_NUM\_STD | 0 | 19063.52 | 50.532 | 174.543 | 574.161 | -473.097 |
| JF\_01\_TIMES\_STD | 0 | 10029.22 | 138.784 | 201.805 | 744.199 | -466.631 |
| JF\_02\_TIMES\_STD | 0 | 19367.13 | 166.936 | 264.505 | 960.451 | -626.579 |
| BIN\_STOP\_STD | 0 | 79.33 | 0.485 | 0.949 | 3.332 | -2.362 |
| ALL\_STOP\_STD | 0 | 19.38 | 0.221 | 0.374 | 1.343 | -0.901 |

按照表3.2中每一变量对应的异常值的上下限来确定异常值，将属于异常值的用户记录删除。在本文的实际数据中，删除异常记录后的总的用户记录数量为31656，其中，正常用户的数量为18259，流失用户的数量为13397。

## 3.3 logistic 回归模型

### 3.3.1 logistic回归模型的原理

Logistic回归模型是为了解决线性概率模型（LPM）的估计和预测中存在的诸多问题而出现的，二项Logistic回归模型（binomial logistic regression model）是一种分类模型，由条件概率分布P(Y|X)表示，形式为参数化的logistic分布，这里随机变量X的取值为实数，随机变量Y取值为0或者1，我们通过监督学习的方法来估计模型参数，logistic回归模型的基本原理如下：

假设训练数据集中共有n个用户，包含K个描述用户属性或者消费行为的特征变量Xi = (x1i,x2i,……xki)，(i = 1,2,……n)，xiR，假设yi为二分类变量，代表用户下月是否流失，当用户下月流失时，yi=1，当用户下月依然处于正常状态，则yi=0，由于在实际情况中，用户流失的概率是无法确定的，但是可以假设有一个理论上存在的连续反应变量代表用户下月流失的可能性，的定义如式（3.5）：

 （3.5）

式（3.5）中，中为随机误差项，通常在多元回归分析中假设服从正态分布，但是在这里假设服从logistic分布，logistic分布的概率密度函数为：

 （3.6）

显然其概率分布函数关于Y轴对称，其分布函数为：

 （3.7）

的值域为负无穷到正无穷，当该变量的值跨越一个临界值（通常取0）时，便导致用户下月流失，于是有：

当时，yi=1。

当时，yi=0。

采用概率的形式表示如下式：

 （3.8）

 （3.9）

依据上式可以求出和，结果如式（3.10）和式（3.11）：

 （3.10）

 （3.11）

式（3.10）和式（3.11）中，为待估计的参数，由于logistic回归为非线性模型，因此可以采用最大似然估计来求解，yi为实际观察到的用户流失情况，显然服从0-1分布，即

 （3.12）

式（3.12）中，由于每个用户是相互独立的，因此似然函数可以定义为：

 （3.13）

对该似然函数取对数似然函数可得：

 （3.14）

为了估计能使该似然函数取最大值，将似然函数对分别求偏导数并且令其为0，则可得：

 （3.15）

 （j=1,2，···，k） （3.16）

式（3.15）和式（3.16）称为似然方程（likelihood equations），对于logistic回归而言，似然方程是的非线性函数，所以求解十分困难，通常只能借助计算机采用迭代方法来求解。

Logistic回归模型估计的一些假设条件与OLS回归十分相似。第一，数据必须来自于随机样本；第二，因变量yi被假设为K个自变量的函数；第三，logistic回归对多重共线性（multicollinearity）敏感，自变量之间存在的多重共线性会导致标准误的膨胀。但是logistic回归模型还有一些与OLS回归不同的假设，首先，logistic回归中的因变量只能是分类变量；其次，logistic回归中的因变量和各自变量之间的关系是非线性的。

### 3.3.2 logistic回归在用户流失预警模型中的测试

本章采用IBM SPSS Modelr14.1建立logistic回归模型，由于SPSS Modeler会自动对分类变量实行重新编码，所以我们只需要设定变量的类型即可，不需要再对字符型变量进行重新编码。

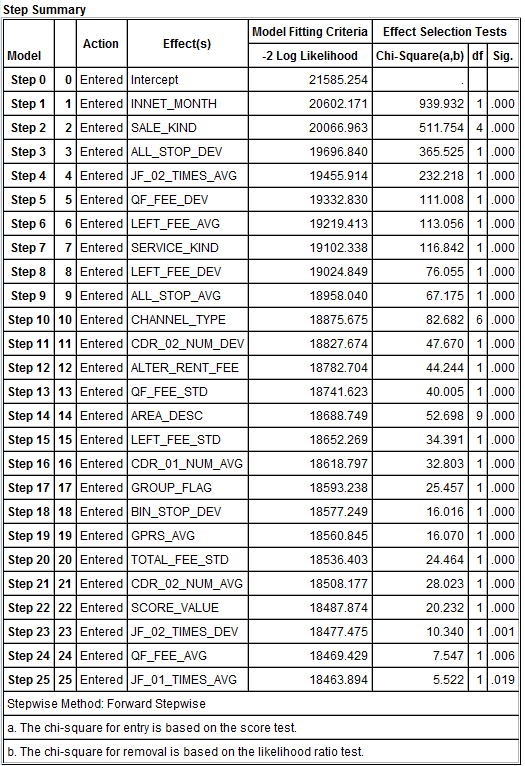
由于logistic回归对多重共线性（multicollinearity）敏感，在使用logistic回归建模时需要对变量进行处理，比较常见的处理方式有：

（1）变量删除。由于自变量之间存在共线性，说明自变量所提供的信息是重叠的，可以删除不重要的自变量以减少重复信息。通常从实际经济分析确定为相对不重要并从偏相关系数检验证实为共线性原因的那些变量中删除。但是如果变量删除不合适将会造成参数估计严重有偏的结果。

（2）采用逐步回归法。逐步回归法是一种常用的消除多重共线性的方法，逐步回归法是逐个引入自变量，引入的条件是该自变量经F检验是显著的。

（3）采用主成分分析法，主成分分析法作为多元统计分析的一种常用方法在处理多变量问题时具有一定的优越性，其降维的优势非常明显。主成分分析法适用于一般的多重共线性问题，当采取主成分提取了新的变量后，往往这些变量间的组内差异小而组间差异大，从而起到了消除共线性的目的。

由于逐步回归法在SPSS Modeler中采用Logistic回归建立模型时非常简单并且有效，因此本文采用逐步回归法来解决输入变量之间存在的多重共线性问题，从迭代结果看，模型经过了25步迭代，Logistic变量选择迭代结果如图3.3所示：



##### 图3.3 logistic回归变量选择迭代结果

本文采用Logistic回归建立用户流失预警模型，具体回归方程如下：

 （3.17）

如果用户的Y\*值大于0，则预测该用户下月将流失。否则该用户为正常用户。将Logistic回归模型的系数与实际情况进行对比发现：（1） [SERVICE\_KIND=2G]的系数为-0.4746 ，表明对比3G用户而言，2G用户的流失率更低，由于运营商在几年以前就已经停止发展2G用户，因此现有的2G用户一般都是在网时间比较长，并且比较忠诚的用户，此类用户流失的可能性比较低，符合实际情况。（2）ALL\_STOP\_DEV的系数为0.5953，表明用户在11月份的双停次数与以往的平均双停次数相比较，11月份的双停次数越多，用户流失的可能性越大；ALL\_STOP\_AVG的系数为0.6497，表明用户在最近半年的平均停机次数越多，则用户流失的可能性越大，这也是符合实际情况的。（3）[GROUP\_FLAG=公众]的系数为0.2353表明相对集团客户而言，公众用户的流失可能性也偏高，这也符合实际情况。所以，从实际角度考虑，logistic回归模型符合实际情况。模型的检验采用训练集和测试集从混淆矩阵角度加以考察，logistic回归模型的效果评价的混淆矩阵如表3.4所示：

#### 表3.4 logistic回归效果评价的混淆矩阵

|  |  |  |  |
| --- | --- | --- | --- |
| 训练数据集 | | 实际值 | |
| 正常 | 流失 |
| 预测值 | 正常 | 7,275 | 1,866 |
| 流失 | 2,782 | 3,953 |
| 测试数据集 | | 实际值 | |
| 正常 | 流失 |
| 预测值 | 正常 | 7,244 | 1,874 |
| 流失 | 2,711 | 3,951 |

从表中可以发现，对训练数据集而言，总的分类准确率为70.72%，精确率为67.9%，召回率为58.7%。对测试数据集而言，总的分类准确率为70.94%，精确率为67.8%，召回率为59.3%。对比训练数据集而言，测试数据集的结果比较稳定，式（3.17）和表3.4表明了在用户流失预警中，使用Logistic回归建立的模型具有很好的稳定性和良好的可解释性。

## 3.4 决策树模型

决策树算法是一种经常用到的分类算法，在分类问题中，表示为基于特征对实例进行分类的过程。决策树模型可以看成是if-then的集合，也可以当成是定义在特征空间和类空间上的条件概率分布，决策树模型是一个树结构。一般地，一颗决策树包含一个根节点，若干内部节点和若干叶节点。每个内部节点表示在一个特征属性上的测试，叶节点对应于决策结果，根节点则包含样本全集。从根节点到每个叶节点的路径对应了一个判定测试序列，决策树模型的主要优点是模型具有可读性并且分类速度快。

### 3.4.1 C5.0模型的原理

ID3算法是由Quinlan在1986年提出来的，经过不断完善形成了具有决策树里程碑意义的C4.5算法，C5.0算法的核心与C4.5相同，只是在执行效率和内存使用方面有所改进。

信息熵（Entropy）是平均信息量的测度指标，从数据分析角度看，极端情况下如果某输入变量只有一个取值，也就是说所有样本在该变量上取相同的值，那么这个输入变量事实上并没有给目标变量的分类预测带来任何有价值的信息，即可以认为该输入变量所提供的信息量为0，因此信息量的多少代表了变量取值差异的程度，信息熵是分类变量取值差异性的测度指标。

假设S是一个训练样本的集合，该集合中每个实例的类编号已知。每个实例为一个元组。存在一个属性用来判定某个实例的类编号，假设S中有m个类，总共s个实例，每个类Ci有si个实例(i＝1,2,3...m)，那么任意一个实例属于类Ci的概率是si / s，那么用来分类的训练样本的信息熵是：

 （3.18）

假设有一个有v个值的属性A{a1,a2,...,av}，则属性A可以将S分成v个子集{S1,S2,...,Sv}，其中Sj包含S中属性A上的值为aj的实例。假设Sj包含类Ci的样本数量为。A的熵定义为：

 （3.19）

使用属性A来划分训练样本S获得的信息增益定义为：

 （3.20）

信息增益反映的是信息消除随机不确定性的程度，信息增益越大，则表明该属性消除样本不确定的程度也就越大。所以可以通过计算S中样本的每个属性的信息增益，来得到一个属性的相关性的排序。

定义为训练样本集S关于属性A的熵，是用来衡量属性A分裂训练样本集S的广度和均匀性。训练样本集在属性A上的取值分布越均匀，的值就越大。的定义如下：

 （3.21）

那么，信息增益率的定义为：

 （3.22）

决策树的生长过程的实质是对训练样本集的不断分组的过程，这涉及到两个方面的问题：第一，如何从众多输入变量中选择一个当前最佳分组变量；第二，如何从分组变量的众多取值中找到一个最佳的分割点。C5.0算法以信息增益率为标准确定最佳分组变量和分割点，因为信息增益率消除了选择那些值较多且均匀分布的属性作为分裂属性的倾向性。如果最佳分组变量是具有K个类别的分类变量，则依据K个取值将样本分成K组，形成树的K个分支，这样处理的好处是在后续的决策树的生长过程中将不会再涉及该变量。如果最佳分组变量是数值型变量，则以MDLP分箱所得到的最小组为界，将小于组限的样本划为一组，大于组限的划为另一组，形成两个分叉。

此外，为了解决决策树可能因为过于依赖训练样本而出现的过拟合现象，C5.0算法利用统计学置信区间的估计方法，直接在训练样本集上估计误差，同时在误差估计的基础上，C5.0将依据“减少-误差”法来判断是否修剪。

### 3.4.2 用户数据的离散化

使用C5.0算法时，需要对连续变量进行离散化处理，变量离散化的统计方法是变量值的分组，常用的分组方法一般可以分为两类：第一类是无指导的数据分组方法，主要包括分位数分组方法、组距分组方法、秩分组方法、均值-标准差分组方法等；第二类是有指导的数据分组方法，主要是指基于最短描述长度原则的熵分组方法。无指导的数据分组方法仅从分组对象（数值型变量）自身出发对变量进行分组，不考虑分组结果对其他变量的影响，而有指导的数据分组方法则考虑分组结果对其他变量的影响，在本文中采用基于最短描述长度的熵分组对联通用户的数据进行离散化。

基于最短描述长度原则，即MDLP（Minimum Description Length Principle）的熵分组是1993年有Fayyad和Irani提出的一种有指导的数据分箱方法，其基本思想是，如果分组后的输入变量对目标变量的取值解释能力显著低于分组之前，那么这样的分组是没有意义的，所以待分组变量应在目标变量的指导下进行分组。基于MDLP的熵分组的核心测度指标是信息熵和信息增益，对于输入变量A（待分组变量），指定一个组限值T，将样本集合S划分为S1和S2，信息熵定义为：

 （3.23）

式中，，为样本集S，S1，S2的样本量。

输入变量A的组限值T带来的信息增益定义为：

 （3.24）

对二分类而言，在组限值T所划分的组S1和S2中，如果S1中样本均为同一类别的样本，S2中样本均为另一类别的样本，那么这个组限值T对输出变量的取值来说是最为理想的，此时的熵最小，信息增益最大。由此可见，信息增益越大，依据组限值T分组输入变量越有意义。按照此种方式，分别计算出组限值T取所有可能值分组后的熵和信息增益，然后选择信息增益最大且有意义的组限值进行分组，此过程可在所得的每个分组中不断重复。

依据MDLP，输入变量A的组限值T带来的信息增益定义为分组所带来的收益，而分组所带来的代价为：

 （3.25）

式（3.25）中，m，m1，m2为样本S，S1，S2中的类别数。

判断分组过程是否应该继续的标准是分组所带来的收益大于分组所带来的代价，即：

 （3.26）

本文采用基于最短描述长度的熵分组对联通用户的数据进行离散化，此种方式也是SPSS Modeler中C5.0算法对连续变量离散化的默认方式，由于连续变量较多，本文只列举出了RENT\_FEE、TOTAL\_FEE\_AVG、CDR\_02\_NUM\_AVG、CDR\_01\_NUM\_AVG这四个字段的分箱结果，具体分箱结果如表3.5所示：

#### 表3.5 连续变量的分箱结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 分箱值 | 下限值 | 上限值 | 记录数量占比 | 记录数量 |
| RENT\_FEE | 1 | 0 | 1 | 40.48% | 6427 |
| 2 | 1 | 5.92 | 10.16% | 1614 |
| 3 | 5.92 | 33.88 | 17.03% | 2705 |
| 4 | 33.88 | 300 | 32.31% | 5130 |
| TOTAL\_FEE\_AVG | 1 | -88.91 | 1 | 13.42% | 2131 |
| 2 | 1 | 5.22 | 6.86% | 1090 |
| 3 | 5.22 | 78.75 | 51.70% | 8209 |

#### 续表3.5 连续变量的分箱结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 分箱值 | 下限值 | 上限值 | 记录数量占比 | 记录数量 |
| TOTAL\_FEE\_AVG | 4 | 78.75 | 356.49 | 28% | 4446 |
| CDR\_02\_NUM\_AVG | 1 | 0 | 1.83 | 15.46% | 2456 |
| 2 | 1.83 | 10.17 | 7.24% | 1150 |
| 3 | 10.17 | 40 | 18.87% | 2996 |
| 4 | 40 | 582 | 58.41% | 9274 |
| CDR\_01\_NUM\_AVG | 1 | 0 | 1.5 | 20.04% | 3183 |
| 2 | 1.5 | 6 | 13.48% | 2141 |
| 3 | 6 | 27.33 | 28.85% | 4581 |
| 4 | 27.33 | 413.33 | 37.61% | 5971 |

### 3.4.3 CART模型的原理

分类回归树（Classification And Regression Tree，CART）是由美国斯坦福大学和加州大学伯克利分校的Breiman等人于1984年提出来的，分类回归树算法同样包括决策树的生长和决策树的修剪两个过程，这点与C5.0算法的思路相同，其主要差别体现在以下几个方面：

（1）C5.0中的输出变量只能是分类变量，而CART中输出变量可以是分类变量也可以是数值型变量。

（2）C5.0可以建立多叉树，而CART只能建立二叉树。

（3）C5.0以信息熵为基础，通过计算信息增益率来确定最佳分组变量和分割点，而CART以Gini系数和方差作为选择依据。

（4）C5.0依据训练样本集，通过近似正态分布确定决策树的标准，而CART则依据测试样本集进行修剪。

分类回归树的生长过程的本质仍然是对训练样本集的反复分组过程，由于在流失预警建模中，目标变量是分类变量（下一个月正常或者流失），因此本文只对分类树作原理上的阐述。在分类树的生长过程中，对每个输入变量都需进行相应的计算以确定最佳分组变量。

（1）对于数值型输入变量。首先，将数据按照升序进行排列，然后从小到大依次以相邻数值的中间值作为组限，将样本分成两组，并计算两组样本输出变量值的差异性。

理想的分组应该尽量使得两组输出变量值的差异性总和达到最小，即“纯度”最大，也就是使两组输出变量值的差异性随着分组而迅速下降，“纯度”迅速增加。

CART采用Gini系数测度输出变量的异质性，Gini系数的定义为：

 （3.27）

式（3.27）中，t为节点；k为目标变量的不同取值数量；P(j|t)是节点t中目标变量取第j类的归一化概率，当节点中所有样本在目标变量上的取值均相同时，目标变量取值的差异性最小，此时Gini系数为0。当节点中目标变量的各类别的样本数量相同时，目标变量取值的差异性最大，此时Gini系数也最大，其值为（1-1/k）。

CART算法采用Gini系数的减少量测度差异性下降的程度，差异性减小量定义为：

 （3.28）

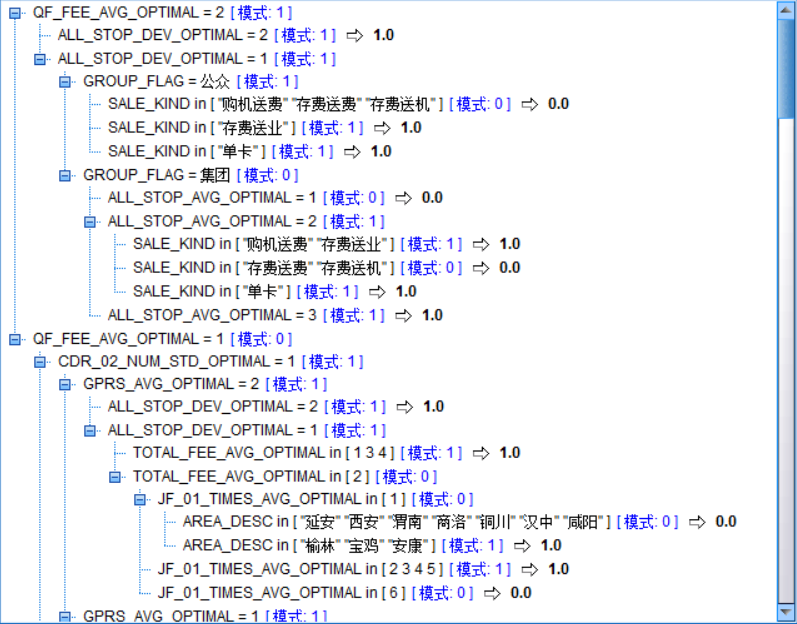
式中，G(t)和N分别为分组前目标变量的Gini系数和样本数量，G(tr)、Nr和G(tl)、Nl分别为两棵子树的Gini系数和两颗子树的样本数量，按照这种计算方法，反复计算便可得到差异性下降最大的分割点，即使达到最大的组限应为当前最佳分割点。

（2）对于分类型输入变量，由于CART只能建立二叉树，对于二类型的输入变量，直接计算Gini系数即可，对于多分类型输入变量，首先需将多类别合并成两个类别，形成超类；然后计算两个超类下输出变量的异质性。

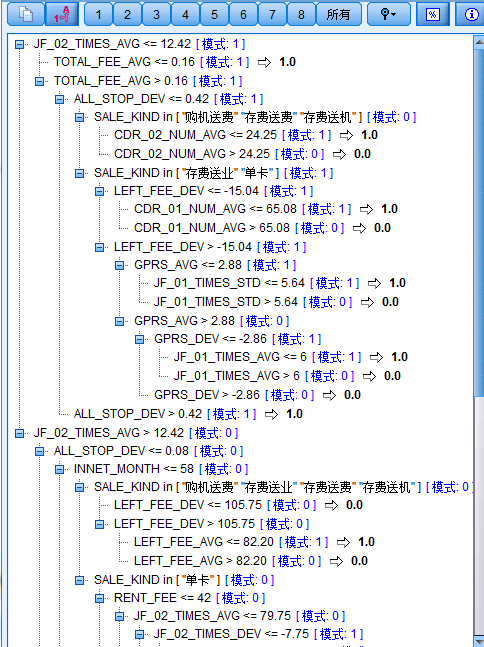
CART的另一核心问题是剪枝策略，CART采用预修剪和后修剪相结合的方式进行剪枝。预修剪的目标是控制决策树的过度生长，可以事先指定一些参数，包括：限制决策树的最大深度，限制决策树中节点的最小样本数量或样本比例，树节点中输出变量的最小异质性减少量等。CART的后修剪策略允许决策树充分生长，然后在此基础上根据一定的规则，剪去决策树中那些不具有一般代表性的叶节点或子树，是一个边修剪边检验的过程，CART采用的后修剪技术称为最小代价复杂性修剪法（Minimal Cost Complexity）。

### 3.4.4 决策树在用户流失预警模型中的测试

本章采用IBM SPSS Modelr14.1利用C5.0和CART决策树算法分别建立用户流失预警模型，C5.0算法和CART算法生成的规则集（判断用户下月即将流失的规则集，如果用户满足以下规则集，则判断该用户为流失用户，否则为正常用户）分别如图3.4和图3.5。

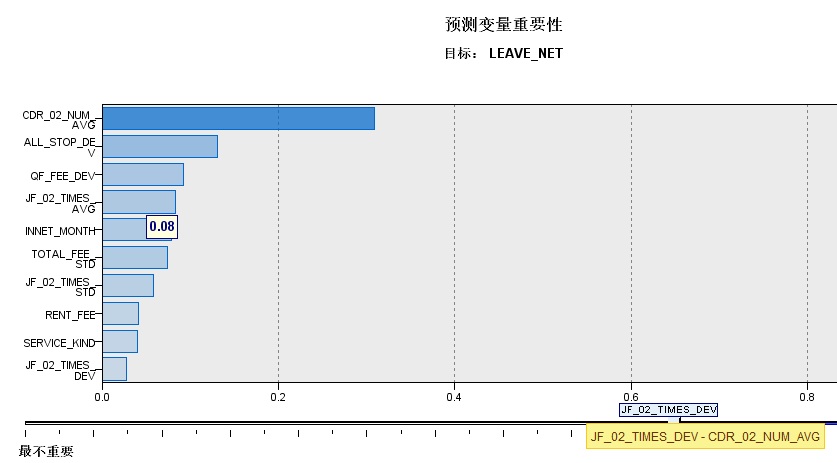


##### 图3.4 C5.0算法建立的流失预警模型规则集



##### 图3.5 CART算法建立的流失预警模型规则集

本文也计算了C5.0决策树模型的预测变量的重要性排序，具体排序结果如图3.6。



##### 图3.6 C5.0决策树预测变量重要性排序

决策树输入变量的重要性测度方法采用了计算输入变量与目标变量相关性统计检验的P值，在此基础上，重要性测度指标是一相对值，第i个输入变量的重要性定义为：，所有进入模型的输入变量的重要性之和为1。在C5.0决策树模型中，最重要的三个输入变量的重要性分别为：CDR\_02\_NUM\_AVG的重要性为0.31，ALL\_STOP\_DEV的重要性为0.13，QF\_FEE\_DEV的重要性为0.09。在C5.0决策树模型中，最重要的变量是平均异网通话次数，异网通话次数越多，转网的可能性越大，这和实际情况一致；其次双停次数离均差反映了用户在11月份的双停次数和最近平均双停次数的差值，停机次数变多，用户感知变差，从而用户流失的可能性也变大；对累计欠费金额而言，累计欠费金额离均差反映了用户在11月的欠费和最近半年的平均欠费金额的变化情况，累计欠费金额增大，说明用户及时缴费的意愿减弱，用户流失的可能性也增大，这些表明，C5.0决策树模型符合实际情况。

C5.0决策树模型的效果评价采用训练集和测试集从混淆矩阵角度加以考察，混淆矩阵如表3.6所示：

#### 表3.6 C5.0决策树模型效果评价的重合矩阵

|  |  |  |  |
| --- | --- | --- | --- |
| 训练数据集 | | 实际值 | |
| 正常 | 流失 |
| 预测值 | 正常 | 7,905 | 1,236 |
| 流失 | 2,491 | 4,244 |
| 测试数据集 | | 实际值 | |
| 正常 | 流失 |
| 预测值 | 正常 | 7,569 | 1,549 |
| 流失 | 2,753 | 3,909 |

从表3.6中可以发现，对训练数据集而言，总的分类准确率为76.52%，精确率为77.4%，召回率为63%。对测试数据集而言，总的分类准确率为72.74%，精确率为71.6%，召回率为58.7%。对比训练数据集而言，测试数据集的结果比较稳定，从而也说明了在用户流失预警中，使用C5.0决策树算法建立的用户流失预警模型具有比较好的命中率和良好的可解释性。

CART决策树模型的检验同样采用训练集和测试集从混淆矩阵角度加以考察，混淆矩阵如表3.7所示：

#### 表3.7 CART决策树模型效果评价的混淆矩阵

|  |  |  |  |
| --- | --- | --- | --- |
| 训练数据集 | | 实际值 | |
| 正常 | 流失 |
| 预测值 | 正常 | 7,710 | 1,431 |
| 流失 | 2,704 | 4,031 |
| 测试数据集 | | 实际值 | |
| 正常 | 流失 |
| 预测值 | 正常 | 7,515 | 1,603 |
| 流失 | 2,805 | 3,857 |

从表3.7中可以发现，对训练数据集而言，总的分类准确率为73.95%，精确率为73.8%，召回率为59.9%。对测试数据集而言，总的分类准确率为72.07%，精确率为70.6%，召回率为57.9%。对比训练数据集而言，测试数据集的结果比较稳定，从而也说明了在用户流失预警中，使用CART决策树算法建立的用户流失预警模型具有比较好的命中率和良好的可解释性。

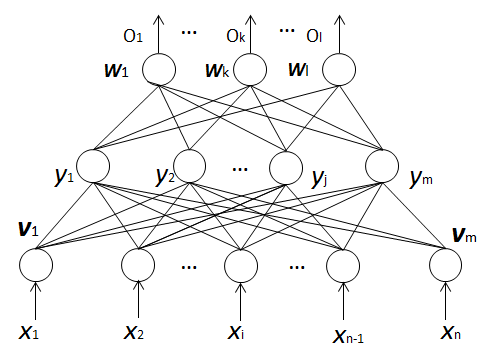
## 3.5 BP神经网络模型

### 3.5.1 BP神经网络模型的原理

BP网络是前馈式神经网络的一种，其核心内容是误差反向传播，BP反向传播网络算法主要思想是把学习过程分为两个阶段：首先是正向传播过程，在此阶段，样本信息从输入层开始逐层经隐节点处理，下层节点的输入为上一层节点的输出，最终样本信息传播到输出层，得到此次传播的输出结果。其次是误差反向传播阶段，在此阶段，计算预测结果与实际结果的差值，此差值记为误差，误差逐层向前传播，传播期间，调整所有网络权值。正向传播和反向传播将会不断进行，从而网络权值也在不断进行调整，直到满足终止条件。

神经网络的层数和每层的节点数决定了网络的复杂程度，其中的关键是隐层的层数和隐层所包含的节点数，隐层层数越多，隐层所包含的节点数越多，网络的复杂程度也越高。对于隐层层数，层数较少，学习时收敛速度较快，但是分类的预测准确率较低。层数较多，分类预测的准确率较高，但是却存在无法收敛的问题。理论上，虽然多层网络能够获得更好的预测结果，但是在实际中，使用两个以上的隐层会使问题的解决变得非常复杂，一般情况下选择一个隐层往往是最合适的，本文使用BP神经网络建立用户流失预警模型也是采用单隐层的网络。

本文以单隐层网络为例进行说明，BP神经网络的结构如图3.7。



##### 图3.7 单隐层BP神经网络结构图

在单隐层BP神经网络中，输入向量记为，隐层输出向量为，输出层输出向量为；假设期望输出向量为，矩阵和矩阵分别表示输入层与隐层和隐层与输出层之间的权值矩阵，向量Vj表示隐层第j个节点对应的权向量；向量Wk表示输出层第k个节点对应的权向量。

输出层输出向量O中的第k个值可由式（3.29）表示。其中，的取值为式（3.30）。

  （3.29）

  （3.30）

隐层输出向量Y中的第j个取值如式（3.31），其中，的取值为式（3.32）。

  （3.31）

  （3.32）

在BP神经网络中，激活函数f通常采用Sigmoid函数，采用Sigmoid函数的优点有：

（1）可以使节点的输出总被约束在[0,1]区间内。如果输出变量为数值型，输出节点给出的是标准化处理后的预测值。如果输出变量为分类型，输出节点给出的是类别为1的概率。

（2）Sigmoid函数很好地体现了在网络权值修正过程中，模型从近似线性到非线性的过程。

当网络输出的结果与期望的输出结果不一致时，则认为存在输出误差，输出误差E通常定义如下：

 （3.33）

式（3.33）中，为第k个输出节点的误差函数，的值为：

 （3.34）

将式（3.29）和式（3.30）代入式（3.33）中，可以得到：

 （3.35）

由式（3.35）可以看出，网络权值调整的目标是使输出误差E达到最小，因此网络权值应沿着误差E曲面下降最快的方向即负梯度方向进行进行调整，由于Sigmoid函数处处可微，的变化量为：

 （3.36）

式（3.36）中，表示学习速率，负号表示负的梯度方向。当采用Sigmoid函数作为激活函数时，的值为：

 （3.37）

因此，的变化量可以表示为：

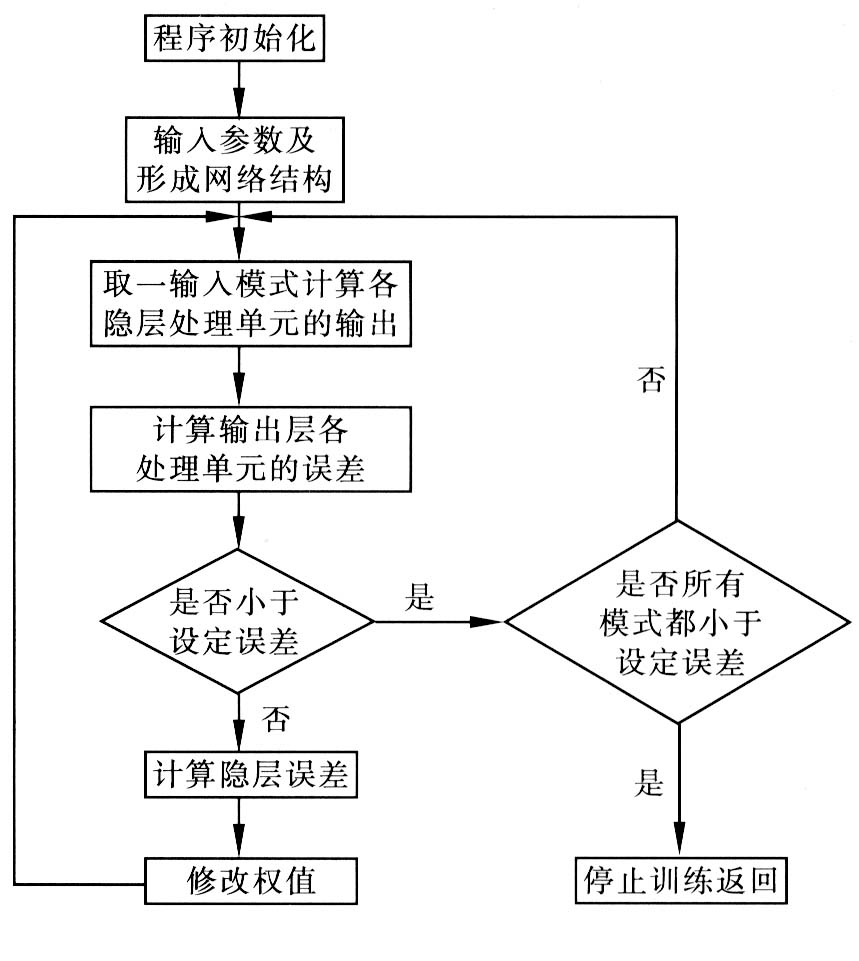
 （3.38）

从式（3.38）中可以发现，如果过大，则网络权值的改变量就较大，这种情况就有可能导致网络的震荡或者发散。反之，如果过小，学习的过程可能就会很长。的取值一般在0.01到0.1之间。

一般情况下，损失函数曲面会存在一些平坦区域，即接近为0，此时模型的学习速度将会很慢，为了解决此种问题，在BP反向传播的网络中，通常会引入冲量项a，即在下一次调整权值的阶段，网络权值将会调整为：

 （3.39）

BP神经网络算法流程如图3.8。



##### 图3.8 BP神经网络算法流程图

### 3.5.2 BP神经网络在用户流失预警模型中的测试

本章采用IBM SPSS Modeler14.1利用神经网络算法建立用户流失预警模型。采用BP神经网络构建分类模型对属性的取值有较严格的要求，输入变量的取值通常要求在0~1之间，否则输入变量的不同数量级别将直接影响权值的确定、从而影响最终的分类预测结果。连续型变量一般采用归一到[0,1]之间。对于分类型输入变量，SPSS Modeler通常采用二进制编码策略来解决此种问题，假设输入变量的不同的取值有k个，则可将该变量转换为n个数值变量，k和n之间的关系满足：

 （3.40）

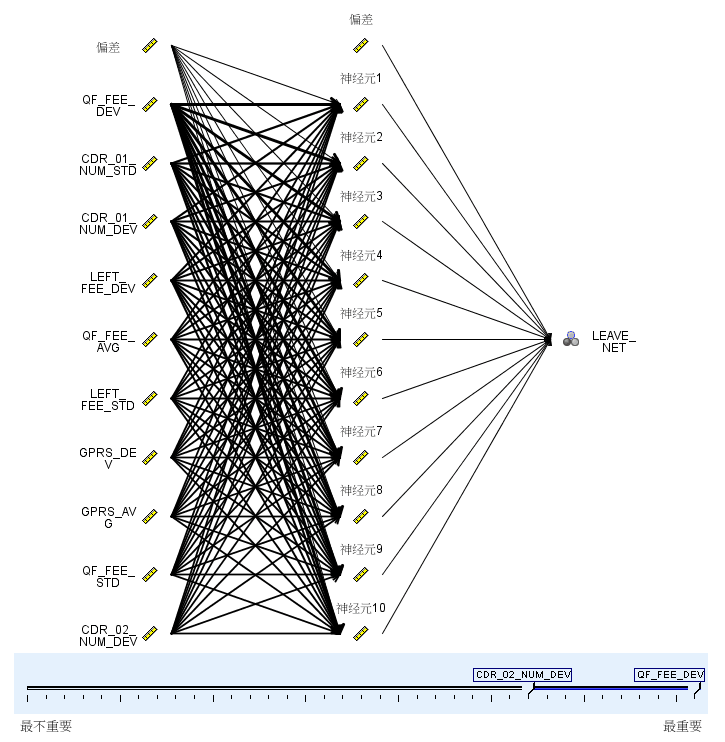
以CHANNEL\_TYPE（渠道类型）为例，CHANNEL\_TYPE为分类型变量，其取值有电话营销、电子渠道、集团直销、社会渠道、社区直销、自有实体渠道和其他总共七个类别，即k=7，此时,也就是说CHANNEL\_TYPE可以编码为000、001、010、011、100、101和110，分别对应话营销、电子渠道、集团直销、社会渠道、社区直销、自有实体渠道和其他这七个类别。

在SPSS Modeler中，BP神经网络学习率的设置策略是：学习率不是一个常数，它会在模型训练中不断进行调整，第t次的训练的学习率与前一次有关，自动调整的计算公式为：

 （3.41）

式（3.41）中和分别表示学习率取值的最大值和最小值，d为衰减量，由于的值总小于1，也就是说学习率处于动态减少过程中。本文中的初始取值为0.3，学习率的最小值为0.01，最大值为0.1，衰减量取值为30，冲量项a取值为0.9。

本文采用BP神经网络建立的用户流失预警模型的结构如图3.9。



##### 图3.9 采用BP神经网络建立流失用户模型的结构

使用BP神经网络建立用户流失预警模型的模型评估采用训练集和测试集从混淆矩阵角度加以考察，混淆矩阵如表3.8所示：

#### 表3.8 BP神经网络效果评价的混淆矩阵

|  |  |  |  |
| --- | --- | --- | --- |
| 训练数据集 | | 实际值 | |
| 正常 | 流失 |
| 预测值 | 正常 | 7,528 | 1,613 |
| 流失 | 2,822 | 3,913 |
| 测试数据集 | | 实际值 | |
| 正常 | 流失 |
| 预测值 | 正常 | 7,474 | 1,644 |
| 流失 | 2,774 | 3,888 |

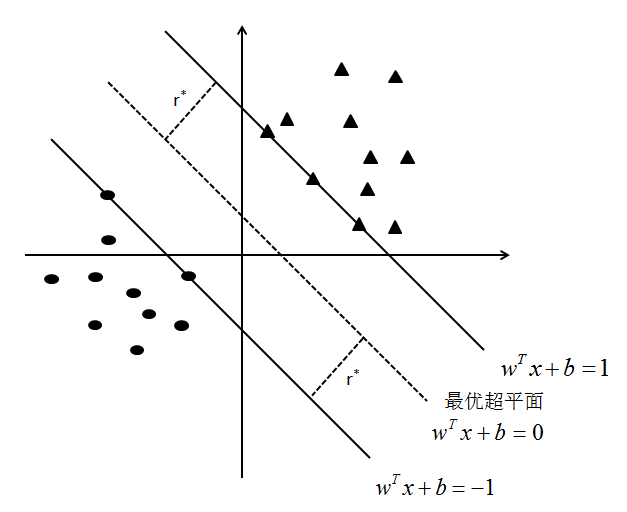
从表中可以发现，对训练数据集而言，总的分类准确率为72.06%，精确率为70.8%，召回率为58.1%。对测试数据集而言，总的分类准确率为72%，精确率为70.3%，召回率为58.4%。对比训练数据集而言，测试数据集的总体分类精度比较稳定，尽管一般认为神经网络的稳定性较差，但是在流失预警中出现了模型预测结果较稳定的情况，同时，神经网络的可解释性很差。

## 3.6 支持向量机模型

### 3.6.1 支持向量机模型的原理

支持向量机（SVM）是在统计学习理论基础之上发展起来的一种数据挖掘方法，通常被认为是在所有知名的数据挖掘算法中最健壮、最准确的方法之一，支持向量机机可以分为支持向量分类机（SVC）和支持向量回归机（SVR）。SVM可以从大量训练的数据中选出很小的一部分用于构建模型，并且对维数不敏感。

使用支持向量机算法来建立模型，实质就是在特征空间中找到一个超平面将样本分开，如果样本数据是二分类并且是线性可分的，那么就存在着多个这样的超平面，支持向量机分类的思想是：找到两个相互平行且间距最大，并且能够将属于不同类别的样本点正确分开的边界，位于两边界中间位置的超平面即为最优超平面。图3.10中显示了线性可分条件下支持向量机的最优超平面。



##### 图3.10 支持向量机的最优超平面

在样本空间中，最优超平面可以通过式（3.42）来描述：

 （3.42）

式（3.42）中，W为法向量，决定了超平面的方向，b为截距，决定了超平面的位置。样本点到最优超平面的几何距离定义为：

 （3.43）

使用SVC建立分类器的目的就是要求出参数值w和b。不失一般性，我们将泛函间隔固定为1，那么对于给定的训练数据集有：

 （3.44）

有一些数据点（xi，yi）使得上式成立，它们就是离最优超平面最近的数据点，这些点被称为支持向量。那么支持向量到最优超平面的几何距离就是：

 （3.45）

为了找到间隔最大的超平面，SVC要用w和b最大化r\*，从而目标函数转化为：

   （3.46）

 由于求的最大值相当于求的最小值，所以目标函数等价于：

   （3.47）

因为目标函数是二次的，约束条件是线性的，所以该问题是一个凸二次规划问题。对式（3.47）可以使用拉格朗日乘子法得到其对偶问题，具体来说，就是对式（3.47）的每条约束添加拉格朗日乘子，所构造的拉格朗日函数如下：

 （3.48）

式（3.48）中是对应第i个不等式的拉格朗日乘数。根据拉格朗日对偶性，原始问题的对偶问题是极大极小问题：

 （3.49）

为了得到对偶问题的解，需要先求对w，b的极小，再求对a的极大。

对的w和b求偏导并令其等于0，则可得到：

 （3.50）

 （3.51）

将式（3.50）和式（3.51）代入式（3.48），则可得：

 （3.52）

即对w，b的极小为：

 （3.53）

将式（3.53）代入式（3.49），可得：

 （3.54）

同时，再加上KKT条件：

  （3.55）

式（3.54）可使用通用的二次规划求解，但是，该问题的规模正比于样本数，这在实际中造成很大的开销，通常的做法是采用SMO算法来解决此问题。

在用户流失预警中，训练样本是非线性可分的，由于两类样本相互交叉，超平面无法将所有样本全部正确分开，此时，超平面的确定应该采用“宽松”的策略，此种情况可以对每个样本点引进一个松弛变量，使函数间隔加上松弛变量大于等于1，这样，约束条件变为：

 （3.56）

是总的错划程度的度量，事实上，只要够大，总能满足约束条件，实际应用中应该避免太大，因此对每个松弛变量乘以代价C（C大于1），则原始问题变为如下的凸二次规划问题：

 （3.57）

惩罚参数C较大时，允许的总错划程度较低，分类的总体精度较高，但是过大的惩罚参数可能导致过拟合问题。惩罚参数C较小时，允许的总错划程度高，分类的总体精度低，惩罚参数C起到了平衡目标函数两个部分大小的作用。此时，原始问题的对偶问题是：

 （3.58）

支持向量机分类解决线性不可分问题的核心思想是空间的非线性变换，即选择一个核函数 κ(⋅,⋅) ，对低纬空间中的线性不可分问题，通过非线性变换，转化为高纬空间中的线性可分问题。常用的核函数有：

（1）线性核函数（Linear Kernel）：

 （3.59）

（2）多项式核函数(Polynomial Kernel)：

 （3.60）

（3）径向基核函数(Radical Basic Function(RBF) Kernel)：

 （3.61）

（4）Sigmoid核函数：

 （3.62）

Tanh函数的形式如下：

 （3.63）

式（3.63）中，为偏差，通常取值为0，增加可提高预测精度，但可能导致过拟合。

### 3.6.2 支持向量机在用户流失预警模型中的测试

本章采用IBM SPSS Modelr14.1利用支持向量机算法建立用户流失预警模型，SPSS Modeler提供了式（3.59）到式（3.62）的四种核函数，经过反复验证，选择径向基核函数作为核函数时效果最优。使用支持向量机求解时，如果采用经典的二次规划方法求解参数，同时求解m个拉格朗日乘子涉及很多次迭代，计算开销很大，因此Modeler采用SMO（Sequential Minimal Optimization）算法求解。其基本思想是每次只更新两个乘子，迭代获得最终解。在模型参数设置中，正则化参数（C）采用默认值设定为10，RBF 伽马设置为0.1，正则化参数（C）的较为合理的值在1~10之间，增加该值将改善模型的精度，但是可能会导致模型过度拟合；RBF伽马的经验值在3/n~6/n之间，n为输入变量的个数，增加伽马可提高预测精度，但同样也可能导致过拟合。

使用支持向量机算法建立用户流失预警模型的评估依然采用训练集和测试集从混淆矩阵角度加以考察，混淆矩阵如表3.9。

#### 表3.9 支持向量机效果评价的混淆矩阵

|  |  |  |  |
| --- | --- | --- | --- |
| 训练数据集 | | 实际值 | |
| 正常 | 流失 |
| 预测值 | 正常 | 7,659 | 1,482 |
| 流失 | 2,385 | 4,350 |
| 测试数据集 | | 实际值 | |
| 正常 | 流失 |
| 预测值 | 正常 | 6,132 | 2,986 |
| 流失 | 1,656 | 5,006 |

从表3.9中可以发现，对训练数据集而言，总的分类准确率为75.64%，精确率为74.6%，召回率为64.6%。对测试数据集而言，总的分类准确率为70.58%，精确率为62.6%，召回率为75.1%。对比训练数据集而言，测试数据集的总体分类精度出现明显的下降，说明支持向量机在用户流失预警中应用时的稳定性较差，同时，支持向量机的固有特点使得支持向量机不具备可解释性。

## 3.7 流失预警单一模型的比较

将本章采用Logistic回归、C5.0、CART、神经网络和支持向量机建立的流失预警模型的测试结果进行比较，结果表3.10所示：

#### 表3.10 流失预警模型效果比较

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 分类器 | 数据集类型 | 准确率 | 精确率 | 召回率 |
| Logistic回归 | 训练数据集 | 70.72% | 67.93% | 58.69% |
| 测试数据集 | 70.94% | 67.83% | 59.31% |
| C5.0 | 训练数据集 | 76.52% | 77.45% | 63.01% |
| 测试数据集 | 72.74% | 71.62% | 58.68% |
| CART | 训练数据集 | 73.95% | 73.80% | 59.85% |
| 测试数据集 | 72.07% | 70.64% | 57.90% |
| BP神经网络 | 训练数据集 | 72.06% | 70.81% | 58.10% |
| 测试数据集 | 72.00% | 70.28% | 58.36% |
| 支持向量机 | 训练数据集 | 75.64% | 74.59% | 64.59% |
| 测试数据集 | 70.58% | 62.64% | 75.14% |

一般而言，分类模型的主要性能评价指标有三个：模型的分类精度、模型的稳健型和模型的可解释性，但是在流失预警的实际应用中，评价用户流失预警模型的性能主要有五个标准：模型的分类精度，模型的稳定度、流失用户的命中率（精确率）、流失用户的覆盖率（召回率）和模型的可解释程度，在本文的研究中，需要重点关注的流失用户的命中率和流失用户的覆盖率和模型的稳定度，而模型的可解释性其次。

从表3.10中的数据可以发现，Logistic回归和BP神经网络的稳定性最好，支持向量机对流失用户的覆盖率最高，而C5.0和CART算法具有最好的流失用户的命中率，从本文流失预警的核心指标上考虑，没有哪一种算法完全优于另外的算法，也没有一种算法在所有指标上而言比其它所有的算法都差。而这些也成为本文构造组合流失预警模型的理论和现实依据。

# 第四章 流失预警组合模型的构建与测试

将组合模型应用于用户流失预警模型有以下几个方面的理由。第一，组合模型能够为用户流失预警模型提供更加广阔的发展空间。目前用户流失预警在银行业、电信业等行业都得到了比较充分的研究，理论日趋完善。改进空间越来越小；第二，组合模型能够为流失预警模型的准确性和稳健性的优化给出了一种相对简单的解决方案。从上一章我们对单一流失预警模型的研究与实证分析中发现，现有的流失预警模型各有自己的优缺点。

本文所构建的用户流失预警组合模型采用的是基于二步聚类的组合分类器模型。组合分类器的构成主要包括两个方面：一方面是基分类器的构造；另一个方面是基分类器的组合。构建具有较大差异性的基分类器的方法有：第一，使用差异性较大的训练数据集应用于同一分类算法构建基分类器；第二，使用同一训练数据集应用于不同分类算法构建基分类器。本文从两个方面均加以考虑，首先使用二步聚类算法将训练样本划分为若干个类别的样本集，对所形成的样本集采用随机抽样的方式选取若干样本集构成训练样本子集，然后将样本子集应用于分类算法从而建立基分类器。

## 4.1 训练样本子集的选取

训练样本子集的构建采用的是两步聚类（two step clustering）算法来完成。两步聚类算法特别适用大型数据集的聚类研究，并且有效克服了K-means方法存在的问题，主要特点表现在：第一，两步聚类算法既可以处理数值型变量，也可同时处理分类型变量；第二，两步聚类可以根据一定的准则自动选择聚类数目；第三，两步聚类能够诊断样本中的离群点和噪声数据；第四，可以处理超大样本量的数据；第五，通过两步实现数据的聚类。

### 4.1.1 两步聚类

两步聚类是一种分层聚类方法，是对BIRCH算法的改进。在两步聚类算法中，聚类特征树中的每个节点只存储聚类过程计算距离的充分统计量。树节点j的充分统计量包括，其中为节点的样本量，为数值型变量值的总和，为数值型变量值的平方和，为分类型变量的样本量。如果第j类和第s类合并，则合并后的充分统计量为：

 （4.1）

两步聚类采用距离测度样本或类的“亲疏程度”，并依据距离确定类的划分，如果聚类变量均为数值型变量，则仍采用欧氏距离，否则采用对数似然距离。两步聚类的对数似然距离设计源于概率聚类的表示方式，假设有K个聚类变量，分别为，其中m个数值型聚类变量和n个分类型聚类变量，且分类变量都服从多项分布，数值型聚类变量服从正态分布（实际上，在两步聚类中，独立性假设和两类变量分布状态假设的违反具有相当强的鲁棒性），如果聚成J类，则对数似然函数定义为：

 （4.2）

式（4.2）中，为似然函数，为第j类样本的集合，为第j类的参数向量。包含若干类的样本，其对数似然是各类样本的对数似然之和。同理，一个由M个子类组成的子类，其对数似然等于M个子类的对数似然之和。

如果第j类和第s类合并，两者合并后的类记为<j，s>，则它们的距离定义为两类合并之前的对数似然与合并之后的对数似然之差，即对数似然距离，具体如下：

 （4.3）

式（4.3）中，，和分别为第k个数值型变量的总方差和在j类中的方差，的作用是为了解决第j类方差可能为0而导致对数无法计算的问题，为第j类的样本量，为在第j类中第k个分量取第l个类别的样本量，第k个分类型变量有个类别。

对数似然函数反映了类内部变量取值的总体差异，且数值型变量以方差测量，分类型变量以信息熵测量。

由于，因此当第j类和第s类合并后，大于0，比较小说明第j类和第s类合并将不会引起类内部的差异性显著增加，当小于某个阈值时，第j类和第s类可以合并，否则，不应合并。

两步聚类需要经过预聚类和聚类这两步实现数据的聚类。预聚类采用“序贯”方式，即数据逐条读入并逐条处理，具体过程为：

（1）建立聚类特征树，其充分统计量存储在特征树的根节点中。

（2）读入一个用户数据，从根节点开始，依据充分统计量使用对数似然距离寻找与该数据最近的中间节点，然后移动到该中间节点并重复该处理直到移动到一个最靠近的叶节点。

（3）如果该用户数据和最近的叶节点的对数似然距离小于阈值T，则该数据被安置在该叶节点中。然后判断该叶节点的样本量是否满足大于阈值B，如果满足，则该节点分裂成两个叶节点，该叶节点变成中间节点。分裂时以距离最远的两点为中心，根据距离最近原则分类。反之，如果该用户数据和最近的叶节点的对数似然距离大于阈值T，该数据将会形成一个新的叶节点。然后更新叶节点及与该叶节点相关的所有中间节点及根节点的充分统计量。

（4）随着CF树的生长，聚类数目不断增加，当叶节点个数达到允许的最大聚类数目时，如果此时数据还没有全部得到处理，则适应增加阈值重新建树，以得到一颗新的较小的CF树。

（5）重复以上的过程，直到所有的用户数据均被分到某个叶节点为止。

在预聚类过程中，如果某个叶节点中的样本量远小于最大叶节点所包含的样本量，则该叶节点中的样本被视为离群点，在聚类阶段，离群点将不会作为此阶段的输入。第二步聚类阶段将会采用层次聚类方法进行聚类。所谓层次聚类，是指将所有的样本点自底向上合并组成一棵树或者是自顶向下分裂成一棵树的过程，在两步聚类算法的聚类阶段，采用的是自底向上合并叶节点的方式。对于N个叶节点，需要进行N-1次迭代，每次迭代过程是计算两两叶节点之间的对数似然距离，并依据距离最小原则，将距离最小的两个叶节点进行合并。在两步聚类中，叶节点数目也就是聚类数目的确定采用的两阶段策略：第一阶段仅给出一个初略的估计；第二阶段给出一个恰当的最终聚类数目，且两个阶段的具体判定标准也不一样。

第一阶段采用的那是贝叶斯信息准则（Bayesian Information Criterion，BIC）作为判断准则，假设聚类数目为J，则有：

 （4.4）

BIC准则第一项反映的是J类对数似然的总和，是对类内差异性的总度量；第二项是一个模型复杂度的惩罚项，当样本确定后，J越大该项值也越大。

合适的聚类数目应是BIC取最小值时的聚类数目。如果所有叶节点合并为一个大类，此时BIC的第一项最大，第二项最小。当聚类数目增加时，第一项开始减小，第二项开始增大，在开始阶段，BIC减小的幅度大于增加的幅度，因此BIC下降，当聚类数据增加到J时，BIC减小的幅度将会小于增加的幅度，此时的J即为最佳的聚类数目。

第二价段是对第一阶段粗略估计值J的修正，依据对数似然距离，在类中选择一个恰当值，不考虑模型的复杂度。所采用的计算方法是：

 （4.5）

式（4.5）中，是聚类数目为J时的两两类间对数似然距离的最小值。是类合并过程中类间差异性最小值变化的相对指标，是一个大于1的数，如果值较大，则说明相对于J+1类J类较合理，不应当继续合并。

### 4.1.2 样本子集选取的测试

本章以上一章所选取的训练数据集为本章样本子集选取的训练数据集，采用IBM SPSS Modelr14.1利用两步聚类算法将用户分类，从而构造样本训练子集。所使用的训练数据集为15876，其中包含9141条正常用户的记录和6735条流失用户的记录。

SPSS Modeler利用BIC的变化量dBIC和变化率在聚类数目确定的第一阶段来确定聚类数目，即：

 （4.6）

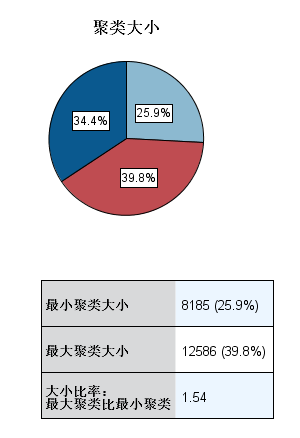
 （4.7）

开始时，如果小于0，则聚类数目为1，后续算法不再继续执行，否则，依次找到取最小值（SPSS Modeler规定应小于0.04）即BIC减小幅度最小时的J为聚类数目的粗略估计。

在确定聚类数目的第二阶段，依次计算、直至，找到其中的最大值和最小值，SPSS Modeler规定，如果最大值是次大值的1.15倍以上，则最大值所对应的J为最终聚类数。

两步聚类结果的评价采用轮廓系数来衡量，轮廓系数的值处于-1~1之间，值越大，表示聚类效果越好。

本文采用两步聚类对训练数据集进行聚类，聚类结果将用户分成了三类，用户类别的占比如图4.1所示。



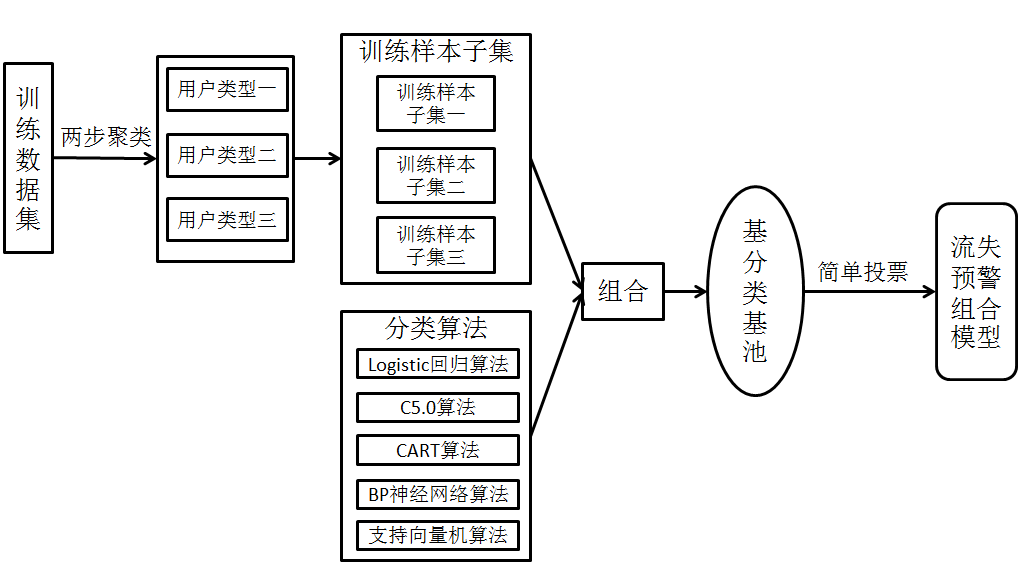
##### 图4.1 采用两步聚类的聚类结果

图4.1的聚类结果显示，轮廓系数为0.3，最大聚类的样本数量是最小聚类样本数量的1.54倍，样本在所聚成的三个类别中分布比较均匀。

## 4.2 用户流失预警组合模型的构建

由于不同分类算法在原理上存在很大的差异性，采用不同的分类算法构成基分类器很好地满足了组合分类器中对基分类器选择的差异性的要求。同时，本文也从训练数据的差异性角度考虑，将同一分类算法应用于不同的具有较大差异性的训练数据集，采用此种方式构成基分类器。

本文构建用户流失预警组合模型的流程如图4.2所示：



##### 图4.2 流失预警组合模型的构建流程

图4.2中，用户类型为采用两步聚类算法所形成的三个类别，上一章对单一分类器的实证结果表明，在通信行业用户流失预警中每一种算法都有其优劣，但是总体分类精度相差不大，因此，在本章中将Logistic回归、C5.0、CART，BP神经网络和支持向量机均作为基分类算法，在样本数据和基分类器的组合中，如果采用两步聚类算法所形成的的单一单一类型的用户样本作为训练数据集建立模型，由于该类别中的样本量比较少，结果出现总体分类精度出现非常明显的下降，因此本文对训练样本子集的选取采用的方式是从两步聚类所形成的的3个类别中随机抽取其中的2个类型的用户数据，为了使构成的每一训练集都有明显的差异，并且包含所有的数据，对训练样本子集的选择有三种：用户类型一和用户类型二、用户类型一和用户类型三、用户类型二和用户类型三。采用此种方式（3种用户类型的样本加上5种分类算法），最终将会形成15种基分类器。

## 4.3 用户流失预警组合模型的测试

### 4.3.1 基分类器差异性的测试

本文对基分类差异性的测试从三个方面进行考虑：Q统计量、不一致度量、协方差度量。由于本文的基分类池由15个基分类构成，如果采用两两进行比较，将会有105个比较结果，但是总体而言，本文对基分类器的构成采用了两种方式：一种是具有差异性的训练样本子集和同一分类算法；第二种是采用了相同训练样本子集和不同分类算法。因此本文仅从这两种情况下选择有代表性的两类加以测试，一类是采用C5.0算法和不同的训练样本子集所构成的3个基分类器，另一类是采用5种分类算法（Logistic回归、C5.0、CART、BP神经网络和SVM）应用于由用户类型一和用户类型二所构成的训练样本子集。结果分别如表4.1和表4.2所示：

#### 表4.1 C5.0算法应用于不同的训练样本子集的差异性测试表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 基分类器1 | 基分类器2 | Q统计量 | 不一致度量 | 协方差度量 |
| c50\_1 | c50\_2 | 0.91 | 0.17 | 0.62 |
| c50\_1 | c50\_3 | 0.84 | 0.21 | 0.53 |
| c50\_2 | c50\_3 | 0.83 | 0.22 | 0.51 |

#### 表4.2 不同算法应用于同一训练样本子集的差异性测试表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 基分类器1 | 基分类器2 | Q统计量 | 不一致度量 | 协方差度量 |
| c50\_1 | cart\_1 | 0.95 | 0.14 | 0.71 |
| c50\_1 | log\_1 | 0.85 | 0.21 | 0.54 |
| c50\_1 | svm\_1 | 0.89 | 0.19 | 0.59 |
| c50\_1 | network\_1 | 0.85 | 0.21 | 0.55 |
| cart\_1 | log\_1 | 0.87 | 0.20 | 0.57 |
| cart\_1 | svm\_1 | 0.90 | 0.18 | 0.61 |
| cart\_1 | network\_1 | 0.86 | 0.21 | 0.55 |
| log\_1 | svm\_1 | 0.96 | 0.11 | 0.75 |
| log\_1 | network\_1 | 0.96 | 0.12 | 0.74 |
| svm\_1 | network\_1 | 0.95 | 0.13 | 0.71 |

从表4.1和表4.2中可以发现采用两种方式的Q统计量的值都非常高，主要原因是基分类器对训练样本的分类准确率都达到了70%以上，所以出现此种现象。同时从数据中还可以发现，两种方式所构成的基分类器的Q统计量、不一致度量和协方差度量的值都非常接近，这说明了本文所采用的采用两步聚类算法来构成有差异性的训练样本子集从而构造有差异性的基分类器是一种切实可行的方法。

### 4.3.2 用户流失预警组合模型准确率的测试

由于用户流失预警组合模型的研究重点是流失用户，因此本文将流失用户作为正例，正常用户作为负例，用户流失预警组合模型在训练集和测试集上的测试结果如表4.3所示：

#### 表4.3 流失预警组合模型的测试结果表

|  |  |  |  |
| --- | --- | --- | --- |
| 测试数据集 | | 实际值 | |
| 正常 | 流失 |
| 预测值 | 正常 | 7694 | 1424 |
| 流失 | 2752 | 3910 |
| 训练数据集 | | 实际值 | |
| 正常 | 流失 |
| 预测值 | 正常 | 7788 | 1353 |
| 流失 | 2648 | 4087 |

表4.3中的数据表明，在训练数据集和测试数据集上，用户流失预警组合模型都显示出良好的拟合结果，不管是在总的准确率上，还是在对流失用户的命中率和覆盖率上，流失预警组合模型都体现除了较优的性能。

采用单一分类算法（Logistic回归、C5.0、CART、BP神经网络和支持向量机）所构建的用户流失预警模型和流失预警组合模型在准确率、平均准确率、精确率、召回率和F1\_SCORE的测试结果对比如表4.4所示：

#### 表4.4 流失预警单一模型与组合模型的测试结果比较

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 分类器 | 数据集类型 | 准确率 | 平均准确率 | 精确率 | 召回率 | F1\_SCORE |
| Logistic回归 | 训练数据集 | 70.72% | 69.14% | 67.93% | 58.69% | 62.98% |
| 测试数据集 | 70.94% | 69.38% | 67.83% | 59.31% | 63.28% |
| C5.0 | 训练数据集 | 76.52% | 74.75% | 77.45% | 63.01% | 69.49% |
| 测试数据集 | 72.74% | 70.84% | 71.62% | 58.68% | 64.50% |
| CART | 训练数据集 | 73.95% | 72.10% | 73.80% | 59.85% | 66.10% |
| 测试数据集 | 72.07% | 70.16% | 70.64% | 57.90% | 63.64% |
| BP神经网络 | 训练数据集 | 72.06% | 70.23% | 70.81% | 58.10% | 63.83% |
| 测试数据集 | 72.00% | 70.17% | 70.28% | 58.36% | 63.77% |
| 支持向量机 | 训练数据集 | 75.64% | 74.19% | 74.59% | 64.59% | 69.23% |

#### 续表4.4 流失预警单一模型与组合模型的测试结果比较

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 分类器 | 数据集类型 | 准确率 | 平均准确率 | 精确率 | 召回率 | F1\_SCORE |
|  | 测试数据集 | 70.58% | 71.20% | 62.64% | 75.14% | 68.32% |
| 组合模型 | 训练数据集 | 74.80% | 72.94% | 75.13% | 60.68% | 67.14% |
| 测试数据集 | 73.54% | 71.54% | 73.30% | 58.69% | 65.19% |

从表4.4中的数据可以看出，在测试数据集上，组合模型在准确率、平均准确率和精确率上优于所有的单一模型，不管是统计模型还是非统计模型。除支持向量机外，组合模型在召回率和F1\_SCORE上要优于所有其他模型，但是使用支持向量机所建立的单一模型在召回率上出现很大的异常，在测试数据集上的召回率居然比训练数据要高出接近10%，这说明在电信行业使用支持向量机算法构建用户流失预警模型具有较大的不稳定性。

### 4.3.3 用户流失预警组合模型稳定性的测试

对于模型稳定性的测试与比较，通常可以采用模型评估准确率的变化大小来区别，由于本文的研究重点是流失用户的预测，因此本文将从准确率的稳定度、精确率的稳定度和召回率的稳定度三个方面来加以考虑，本文采用训练数据集和测试数据集的准确率的标准差来衡量准确率的稳定度，精确率的标准差来衡量精确率的稳定度，召回率的标准差来衡量精确率的稳定度，准确率的稳定度具体的计算公式如下：

 （4.8）

式中，为稳定度值，为训练数据集的准确率，为测试数据集的准确率，为和的均值。精确率和召回率的稳定度同样采用式（4.8）进行计算。

用户流失预警组合模型与采用单一分类算法所构建的用户流失预警模型的稳定度测试对比结果如表4.5所示：

#### 表4.5 组合模型与单一模型稳定度测试对比表

|  |  |  |  |
| --- | --- | --- | --- |
| 分类器 | 准确率稳定度 | 精确率稳定度 | 召回率稳定度 |
| Logistic回归 | 0.001563619 | 0.000737571 | 0.004335426 |
| C5.0 | 0.026775806 | 0.041193316 | 0.03067452 |
| CART | 0.013353643 | 0.022343018 | 0.013830974 |
| BP神经网络 | 0.00043994 | 0.003738596 | 0.001848181 |

#### 续表4.5 组合模型与单一模型稳定度测试对比表

|  |  |  |  |
| --- | --- | --- | --- |
| 分类器 | 准确率稳定度 | 精确率稳定度 | 召回率稳定度 |
| 支持向量机 | 0.035775804 | 0.084505198 | 0.07463248 |
| 组合模型 | 0.008925924 | 0.012907099 | 0.01408497 |

从表4.5中数据可以看出，采用Logistic回归建立用户流失预警模型在准确率稳定度、精确率稳定度和召回率稳定度上均表现出了最优异的性能，这和普遍认为的在电信行业用户流失预警中各种分类算法的表现一致。支持向量机在稳定度方面变现得最差，这也是在实际中，电信行业很少采用支持向量机应用于用户流失预警的原因。本文所研究的组合模型在稳定度上优于C5.0算法、CART算法和支持向量机，尽管其稳定度要比Logistic回归和BP神经网络要差，但是依然表现出了优异的性能，从稳定性角度考虑，也依然具有很大的实际应用价值。

# 第五章 个性化4G套餐推荐系统的设计与测试

上一章所研究的用户流失预警组合模型的目的是为了提高对流失用户的识别的准确性和稳定性，但是并没有给出对流失用户的维系挽留策略。在运营商内部，用户维系挽留通常属于客户部门的工作内容，而用户的历史数据的分析却属于信息化部门，客服人员只能接触到经分平台上的用户上月的数据，只通过对上月用户的消费数据分析来实施对用户的维系挽留策略具有很大的局限性，以4G套餐推荐而言，用户的当月的消费情况并不能很好地表现用户的消费偏好，本章采用用户在迁转4G的前半年的消费的集中趋势和离散趋势的数据，建立个性化4G套餐推荐系统作为对预流失用户的维系挽留策略，从而和上一章结合起来构成了预流失用户从发现到维系的一整套策略

## 5.1 个性化推荐在4G套餐上的应用

个性化推荐[52][53]研究直到20世纪90年代才被作为一个独立的概念提出来，个性化推荐被认为是解决当前信息超载问题最有效的方法，推荐问题从根本上来说就是评估用户从未看到过的产品，推荐系统通常包括三个部分：一是行为记录模块，该模块的作用是记录用户的偏好行为，对运营商而言，用户的偏好行为表现为用户的消费行为特征，以数据库中表的形式表示，该表记录了用户的消费行为数据；二是模型分析模块，该模块的作用是通过分析用户的行为记录来发现用户的潜在感兴趣的产品和感兴趣的程度，由于本章研究的目的是对用户进行个性化的4G套餐推荐，不涉及到其他的通信类业务，因此，本章的研究不涉及到此模块；三是推荐算法模块，该模块的作用是使用推荐算法从4G产品集合中筛选出用户感兴趣的4G套餐对用户进行推荐。在推荐系统的三个组成部分中，推荐算法模块在推荐系统中起到最为核心的作用。

个性化推荐是根据用户的兴趣特点和购买行为，向用户推荐用户感兴趣的信息和商品。根据推荐算法的不同, 推荐系统可以分为如下几类:协同过滤(collaborative filtering)推荐系统;基于内容(content-based)的推荐系统和混合( hybrid) 推荐系统等。

基于协同过滤的推荐最早被提出并且在实际中得到了广泛应用，其核心思想可以分为两部分：首先是利用用户的历史信息计算用户之间的相似度；然后利用与目标用户相似度较高的用户对其他产品的评价来预测目标用户对特定产品的喜好程度，系统根据这一喜好程度来对目标用户进行推荐。协同过滤推荐系统的最大优点是对推荐对象没有特殊的要求。基于内容的推荐源于信息获取领域，该方法是根据用户已经选择的对象，从推荐对象中选择其他特征相似的对象作为推荐结果。不管是协同过滤推荐还是基于内容的推荐，每种推荐算法都有各自的优缺点，混合推荐系统就是通过采用多种不同的推荐算法进行组合，从而达到充分利用各种推荐算法的优势，进而提高推荐系统的准确率。

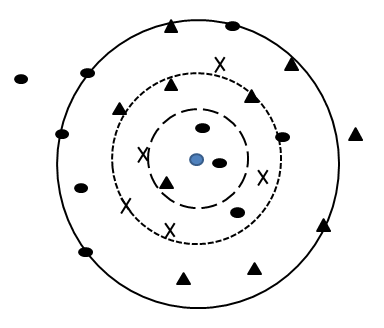
个性化的4G套餐推荐作为一种对流失用户的维系挽留策略，从运营商和用户角度考虑有三种推荐方法：一是采用使用户具有最低出账金额的4G套餐推荐，此种方式对用户是最优的，但是对运营商而言是最不利的，同时，选择什么样的4G套餐会使得用户在未来的一段时间刚好满足用户的需求是未知的；第二种是采用使用户在当前出账金额不变的前提下具有最佳套餐档位的4G套餐推荐，此种方法对运营商而言是最优的，但是由于同档位的4G套餐资费要远高于3G和2G套餐，尽管用户对未来的消费情况未知，但是有一个基于当前消费情况的基本的预估，所以用户肯定不会选择所推荐的套餐；第三种是采用向流失用户推荐和该用户具有最相似消费行为的用户所选择的4G套餐，此种方式避免了前两种方式必然会对用户和运营商之一产生最不利影响的方法，因此，本文采取第三种方式对流失用户进行个性化的4G套餐推荐，即将具有和用户最相似的转4G用户所选择的4G套餐推荐给用户。

个性化推荐技术成熟于电子商务领域，对电子商务领域而言，商品种类繁多，但是对通信行业的4G套餐而言，种类非常有限，对普通用户而言，不管是使用哪一个运营商所提供的服务，主流的4G套餐至多只有几十种，由于4G套餐种类有限，并且推荐信息一般以短信方式下发，TOP-N推荐的方式在此处显得并不适合，结合运营商的实际情况，本文采用的推荐方式是只向用户推荐一种套餐。采用基于用户的协同过滤算法来对用户进行4G套餐的推荐，最核心的思想就是计算用户之间的相似度来寻找即将流失的用户的最近邻。常用的用户的相似度的计算方法有皮尔森相关系数、余弦相似度和修正的余弦相似度等度量方法，假设用户的通信消费数据用向量表示，表示用户在消费特征上的消费值，用户的通信消费数据用向量表示，那么用户和用户的相似性即可由向量和向量之间的相似度表示。由于消费数据向量和用户所选择的4G套餐直接相关，甚至直接决定了4G套餐的可选范围，因此，在协同过滤推荐算法中，常用的相似度衡量方式在此处也不合适，本文所采用的方法是先将用户的消费数据标准化，然后采用欧氏距离计算用户之间的距离，使用距离来衡量用户之间的相似度，距离越近，用户相似度越高。

综合以上分析可以得出，使用基于用户的协同过滤算法向用户所推荐的4G套餐等价于采用以用户所选4G套餐作为目标向量的KNN分类算法对用户进行的分类预测。对本文的研究而言，两者具有一致性。

## 5.2 基于KNN算法的个性化推荐

K最近邻(k-Nearest Neighbor，KNN)分类算法是一种常用的监督分类方法，该算法是“懒惰学习”的一种，即在训练阶段仅仅是把样本保存起来，训练时间的开销为零，当进行预测时再进行处理。KNN算法的工作机制非常简单：给定一个测试样本，按照所选取的距离度量方法在特征空间中找到与测试样本距离最近的k个样本，然后基于这k个样本来对测试样本进行预测。对分类问题而言，一般采用的是简单投票法（还可以采用投票等其他方式），也就是选择k个样本中出现最多的类别标记作为对测试样本的分类预测。使用KNN算法寻找最近邻集的示意图如图5.1.



##### 图5.1 使用KNN算法寻找最近邻集示意图

KNN算法的具体执行步骤如下：

（1）构建训练样本集合。假设有一组已迁转4G的用户样本数据，集合定义为T。其中，T由N个用户的样本数据组成，其数学表示式为：

 （5.1）

式（5.1）中为第i个用户的特征向量，为第i个用户所签转的4G套餐。

（2）设定K值。K值的选取没有一个统一的方法，通常采用的方法是首先确定一个初始值，然后根据测试的结果不断对K值进行调整，最终达到最优。

（3）在训练样本集T中选出与待测样本最近的K个用户的样本，假定样本点，样本点之间的距离一般用欧氏距离来度量。设第i个样本点为，则待测样本与第i个样本点的距离为：

 （5.2）

涵盖这K个样本点的x的领域记为。

（4）在中根据分类决策规则（如多数表决）确定的类别：

 （5.3）

式中，I为指示函数，即当时I为1，否则为0。

KNN分类算法是一种非参数的分类方法，该算法的优点有：第一，简单，易于理解，易于实现，无需估计参数，无需训练；第二，适合对稀有事件进行分类；第三，特别适合于多分类问题(multi-modal,对象具有多个类别标签)；第四，KNN算法最直接地利用了样本之间的关系，减少了特征选择不当的影响，对于一些特征不明显的类别而言，KNN算法更能体现其分类规则独立性的优势。

## 5.3 个性化4G套餐推荐系统的测试

本章选取联通A省分公司在2015年12月份转4G的38294户用户，并且以这些用户在最近半年内（2015年6月份到11月）的数据为基础，将用户分成训练集（用户数量19179）和测试集（用户数量19115），用户所订购套餐通常只和其消费情况有关，在个性化4G套餐推荐系统中，只选取了用户的出账金额、语音、流量和短信的消费属性，具体所选取的字段如表5.1。

#### 表5.1 用户消费属性字段选择

|  |  |
| --- | --- |
| TOTAL\_FEE\_AVG | 平均出账收入 |
| CDR\_02\_NUM\_AVG | 平均异网通话次数 |
| CDR\_01\_NUM\_AVG | 平均同网通话次数 |
| JF\_01\_TIMES\_AVG | 平均本地主叫计费时长 |
| JF\_02\_TIMES\_AVG | 平均本地被叫计费时长 |
| GPRS\_AVG | 平均上网流量 |
| SMS\_NUM\_AVG | 平均短信数量 |
| TOTAL\_FEE\_STD | 出账收入标准差 |
| CDR\_02\_NUM\_STD | 异网通话次数标准差 |
| CDR\_01\_NUM\_STD | 同网通话次数标准差 |
| JF\_01\_TIMES\_STD | 本地主叫时长标准差 |
| JF\_02\_TIMES\_STD | 本地被叫时长标准差 |
| GPRS\_STD | 上网流量标准差 |
| SMS\_NUM\_STD | 短信数量标准差 |

分析此38294户用户所转的4G套餐的情况，所转4G套餐对应的套餐名称和套餐编码以及其对应的用户数量如表5.2所示，由于某些套餐的特殊性，比如用户数量非常少，或者是组合套餐（无法区别具体的套餐内容），或者是校园套餐等，本文不对此部分加以考虑，所以本文实际选择的用户4G套餐迁转数据如表5.2所示。

#### 表5.2 用户4G套餐迁转数据

|  |  |  |
| --- | --- | --- |
| 套餐编码 | 套餐名称 | 用户数量 |
| 1 | 4G本地套餐-36元套餐 | 1757 |
| 2 | 4G本地套餐-56元套餐（陕） | 551 |
| 3 | 4G本地套餐-6元套餐 | 5473 |
| 4 | 4G本地语音套餐-26元套餐（陕） | 4277 |
| 5 | 4G本地语音套餐-36元套餐（陕） | 1993 |
| 6 | 4G本地语音套餐-46元套餐（陕） | 1060 |
| 8 | 4G全国套餐-106元套餐 | 8571 |
| 9 | 4G全国套餐-136元套餐 | 4710 |
| 10 | 4G全国套餐-166元套餐 | 2105 |
| 11 | 4G全国套餐-196元套餐 | 1143 |
| 12 | 4G全国套餐-296元套餐 | 533 |
| 13 | 4G全国套餐-396元套餐 | 283 |
| 14 | 4G全国套餐-596元套餐 | 133 |
| 15 | 4G全国套餐-76元套餐 | 5705 |

使用KNN算法寻找和测试用户具有最相似消费行为的用户，属性变量的选择具有关键性的影响，本文对属性变量的选择采用的是从属性变量与目标变量的相关性的角度来进行，由于属性变量均为数值型变量，目标变量为4G套餐为分类型变量，本文采用计算方差分析中的F检验统计量和对应的1-概率P值来进行判断，判断属性变量很重要、中等重要和不重要的标准值为1-概率P值大于0.95为重要，在0.9~0.95之间为中等重要，小于0.9为不重要，属性变量的重要性检测结果如图5.2。



##### 图5.2 属性变量的重要性检测结果

从图5.2中可以发现，所有14个变量均为重要变量，也就是说使用KNN算法时，这些属性变量均应作为输入变量纳入模型考虑。

KNN算法的核心是K值的选取和距离的测度，本文采用标准化的欧氏距离作为距离的测度方式，K值的选取对推荐的准确度有非常大的影响。通常而言，K值太小，推荐效果容易受到噪声的影响，K值过大，所提取的邻近子集中由于包含了过多的错误点也对推荐准确度产生了影响，通常而言，K值一般取基数值，本文分析了K分别取1、3、5、7、9、11、13和15时推荐的准确度进行测试，测试结果如表5.3。

#### 表5.3 K值选择测试结果表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K值 | 训练集 | | | 测试集 | | |
| k=1 | 正确 | 11,829 | 61.68% | 正确 | 11,786 | 61.66% |
| 错误 | 7,350 | 38.32% | 错误 | 7,329 | 38.34% |
| k=3 | 正确 | 9,633 | 50.23% | 正确 | 9,587 | 50.15% |
| 错误 | 9,546 | 49.77% | 错误 | 9,528 | 49.85% |
| k=5 | 正确 | 8,762 | 45.69% | 正确 | 8,744 | 45.74% |
| 错误 | 10,417 | 54.31% | 错误 | 10,371 | 54.26% |
| k=7 | 正确 | 8,375 | 43.67% | 正确 | 8,300 | 43.42% |
| 错误 | 10,804 | 56.33% | 错误 | 10,815 | 56.58% |
| k=9 | 正确 | 8,072 | 42.09% | 正确 | 7,970 | 41.70% |

#### 续表5.3 K值选择测试结果表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K值 | 训练集 | | | 测试集 | | |
| k=9 | 错误 | 11,107 | 57.91% | 错误 | 11,145 | 58.30% |
| k=11 | 正确 | 7,967 | 41.54% | 正确 | 7,799 | 40.80% |
| 错误 | 11,212 | 58.46% | 错误 | 11,316 | 59.20% |
| k=13 | 正确 | 7,811 | 40.73% | 正确 | 7,628 | 39.91% |
| 错误 | 11,368 | 59.27% | 错误 | 11,487 | 60.09% |
| k=15 | 正确 | 7,736 | 40.34% | 正确 | 7,507 | 39.27% |
| 错误 | 11,443 | 59.66% | 错误 | 11,608 | 60.73% |

以表5.3中的数据为基础，比较K值变化时，测试集和训练集中的用户所选4G套餐和推荐套餐的匹配程度，结果如图5.3。

##### 图5.3 K值和推荐准确率的关系图

从图5.3中可以看出，随着K值的增加，不管是在测试集还是训练集上的推荐的准确性都呈现出下降得趋势，并且KNN算法表现出很稳定的特性。因此，本文采用K=1建立KNN模型。当K等于1时，从每一用户所办理的4G套餐的角度考虑，推荐套餐和用户实际办理的套餐的匹配关系的KNN混淆矩阵如表5.4。

#### 表5.4 KNN混淆矩阵

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 测试集 | 套餐编码 | | 推荐套餐 | | | | | | | | | | | | |
| 1 | 10 | 11 | 12 | 13 | 15 | 2 | 3 | 4 | 5 | 6 | 8 | 9 |
| 实际办理套餐 | 1 | 114 | 28 | 0 | 0 | 0 | 151 | 0 | 147 | 163 | 70 | 0 | 133 | 70 |
| 10 | 0 | 319 | 0 | 0 | 0 | 124 | 0 | 40 | 47 | 0 | 0 | 340 | 187 |
| 11 | 19 | 69 | 137 | 0 | 0 | 53 | 0 | 14 | 19 | 12 | 0 | 151 | 121 |
| 12 | 5 | 26 | 46 | 46 | 0 | 25 | 2 | 6 | 3 | 4 | 5 | 52 | 41 |
| 13 | 2 | 5 | 7 | 10 | 36 | 10 | 0 | 7 | 0 | 2 | 6 | 20 | 25 |
| 14 | 3 | 12 | 6 | 3 | 4 | 7 | 4 | 4 | 0 | 0 | 0 | 12 | 8 |
| 15 | 0 | 0 | 0 | 0 | 0 | 2,141 | 0 | 0 | 0 | 0 | 0 | 703 | 0 |
| 2 | 18 | 14 | 4 | 0 | 0 | 55 | 7 | 26 | 33 | 22 | 6 | 67 | 21 |
| 3 | 0 | 0 | 0 | 0 | 0 | 270 | 0 | 2,274 | 0 | 0 | 0 | 268 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 277 | 0 | 494 | 947 | 0 | 0 | 273 | 90 |
| 5 | 0 | 33 | 0 | 0 | 0 | 177 | 0 | 146 | 165 | 207 | 0 | 198 | 78 |
| 6 | 23 | 23 | 13 | 0 | 0 | 85 | 0 | 51 | 52 | 44 | 41 | 118 | 69 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4,292 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 326 | 0 | 113 | 0 | 0 | 0 | 665 | 1,268 |
| 训练集 | 实际办理套餐 | 1 | 115 | 21 | 0 | 0 | 0 | 133 | 0 | 174 | 152 | 57 | 0 | 161 | 68 |
| 10 | 0 | 322 | 0 | 0 | 0 | 141 | 0 | 45 | 43 | 0 | 0 | 287 | 210 |
| 11 | 17 | 61 | 127 | 0 | 0 | 38 | 0 | 31 | 18 | 5 | 0 | 136 | 115 |
| 12 | 2 | 33 | 27 | 42 | 0 | 21 | 1 | 5 | 7 | 7 | 8 | 51 | 68 |
| 13 | 0 | 12 | 16 | 9 | 48 | 6 | 1 | 5 | 6 | 4 | 4 | 26 | 16 |
| 14 | 0 | 7 | 9 | 8 | 2 | 7 | 1 | 1 | 0 | 4 | 0 | 15 | 16 |
| 15 | 0 | 0 | 0 | 0 | 0 | 2,197 | 0 | 0 | 0 | 0 | 0 | 664 | 0 |
| 2 | 13 | 10 | 6 | 0 | 0 | 41 | 15 | 30 | 33 | 17 | 5 | 75 | 33 |
| 3 | 0 | 0 | 0 | 0 | 0 | 239 | 0 | 2,162 | 0 | 0 | 0 | 260 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 333 | 0 | 515 | 962 | 0 | 0 | 292 | 94 |
| 5 | 0 | 24 | 0 | 0 | 0 | 169 | 0 | 119 | 178 | 212 | 0 | 207 | 80 |
| 6 | 26 | 28 | 7 | 0 | 0 | 90 | 0 | 52 | 74 | 31 | 55 | 115 | 63 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4,279 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 330 | 0 | 112 | 0 | 0 | 0 | 646 | 1,250 |

推荐系统常用的分类准确度的评估方式有：准确率、召回率、准确率提高率、召回率提高率，F1指标和ROC曲线面积，还有覆盖率、多样性和新颖性等，但是这些指标多用于TOP-N推荐，电信行业的个性化4G套餐推荐不属于TOP-N推荐。因此，推荐系统常见的评估指标不适合对本文个性化4G套餐推荐系统的评价，本文除采用上文中所采用的准确率评估外，本文也从4G套餐角度加以评估，由于用户对4G套餐的选择是非均匀的，如果只考虑准确率，可能会出现由于有些4G套餐的用户量比较小从而对此类用户全部产生错误推荐的情况，在本文中，套餐为4G全国套餐-596元套餐的用户的实际选择和推荐套餐全部不一致，为了将此种情况纳入考虑，本文采用平均准确率的方式来评估本文的推荐效果：平均准确率的定义如下：

 （5.4）

式（5.4）中N为4G套餐的数量，表示第i个套餐的用户量，为推荐准确的用户量。

测试数据集的平均准确率为35.55%，训练数据集的平均准确率为36.15%，训练集和测试集的准确率和平均准确率同时也表明模型具有很好地稳定性。说明本文所构造的个性化4G套餐推荐系统取得了比较良好的结果。

由于当前在已知的文献和在运营商实际应用中，都没有对3G或者4G套餐的推荐准确率进行统计，本文第一次采用用户流失预警与个性化推荐相结合的方式来解决了电信行业存量用户中即将流失的用户的识别及其维系策略，对当前运营商的存量用户经营具有非常大的现实意义。

# 第六章 总结

本文以联通A省分公司的用户数据为基础，比较了五种常见分类算法（Logistic回归、CART、C5.0、BP神经网络和支持向量机）在电信行业用户流失预警建模中的效果，并且在此基础上构建了流失预警组合模型，在准确率、精确率、召回率以及稳定性上均取得了较好的结果，同时也对流失用户提出了一种新的维系挽留策略——对流失用户进行个性化的4G套餐推荐，从而提升了用户的感知。

本文所构建的用户流失预警组合模型在准确率、平均准确率和精确率要优于所有单一采用Logistic回归、CART、C5.0、BP神经网络和支持向量机算法建立的用户流失预警模型，召回率和F1\_SCORE也优于除采用支持向量机建立的用户流失预警模型之外的所有模型，同时本文也发现在电信行业用户流失预警中，采用支持向量机建立的模型的稳定性较差。结合稳定性考虑，本文所构建的用户流失预警组合模型要优于所有单一模型。

本文所构建的个性化4G套餐系统在训练集和测试集上的测试结果也表明采用KNN算法向用户推荐的套餐和用户实际选择的套餐的匹配程度较高，说明本文对此部分的研究也取得了良好的效果。

总体而言，本文首先分析了当前通信运营商尤其是中国联通在存量用户流失上所面临的困境，同时分析了通信行业在用户流失预警问题上的研究现状以及实际应用情况，并指出了其不足之处。结合以上现象，本文提出了采用组合算法来构建用户流失预警组合模型，并将此流失预警组合模型和个性化4G套餐推荐系统相结合，形成了一套从流失用户的发现到维系挽留的一整套策略。本文的研究具有以下的特点：

（1）本文不仅给出了流失用户的判别方法，同时也给出了预流失用户的维系策略。

（2） 当前电信行业对用户流失预警的研究在建模数据选择上通常没有考虑数据的可获取性和真实性，这样导致模型在实际应用时具有很大的局限性，并且应用成本非常高昂，本文选取联通某省份公司BSS上的数据，同时放弃了诸如用户年龄、性别、用户性质这样有很大误差或者缺失的数据来建立模型。因此，本文所构建的模型很容易应用于实际，且不需要对原有系统进行大的改造。

（3）本文采用联通某省份公司的数据将通信行业用于用户流失预警的分类方法进行了实证研究，比较了不同分类算法在准确率和稳定度方面的差异。

（4）本文采用两步聚类算法形成差异性较大的训练数据集，从而在此基础上构造基分类器，解决了组合分类器对基分类器差异性的需求。

（5）本文采用KNN算法建立个性化4G套餐推荐系统，该方法具有简单、易用并且准确率较高的特点，不仅为流失用户的维系挽留提供了一种非常有效的手段，并且也对运营商发展4G用户提供了一种良好的策略。

综上所述，本文的研究具有很大的现实意义和参考价值，对当前运营商在用户流失预警和存量用户维系上起到了很好的借鉴作用。

# 参考文献

[1]Poel D V D, Larivière B. Customer attrition analysis for financial services using proportional hazard models[J]. European Journal of Operational Research, 2004, 157(1):196-217.

[2]Maack R, Lück H. An econometric model for resource management in competitive wireless data networks[J]. IEEE Network the Magazine of Global Internetworking, 2004, 18(6):20-26.

[3]骆品亮. 3G背景下电信竞争模型与接入规制政策研究[J]. 产业经济研究, 2006(1):19-25.

[4]周慧, 胡汉辉. 携号转网对于加强电信市场竞争效应的研究[J]. 中国软科学, 2009(1):28-33.

[5]刘玉芹, 胡汉辉. 电信产业链网状化与电信市场竞争[J]. 中国工业经济, 2011(10):130-140.

[6]丁文辉. 从价格竞争到差异化竞争——我国电信市场竞争走向分析[J]. 广东商学院学报, 2006(5):23-27.

[7]Rodriguez S, Shin H. Developing customer churn models for customer relationship management[J]. International Journal of Business Continuity & Risk Management, 2013, 4(4):302-322.

[8]Yu W, Jutla D N, Sivakumar S C. A churn-strategy alignment model for managers in mobile telecom[C]// Communication Networks and Services Research Conference, 2005. Proceedings of the 3rd Annual. IEEE, 2005:48-53.

[9]Yan L, Wolniewicz R H, Dodier R. Predicting Customer Behavior in Telecommunications[J]. Intelligent Systems IEEE, 2004, 19(2):50-58.

[10]Galbi D A. Regulating prices for shifting between service providers[J]. Information Economics & Policy, 2001, 13(2):181-198.

[11]廖里, 余英泽, 吴渝,等. 数据挖掘和数据仓库及其在电信业中的应用[J]. 重庆邮电大学学报:自然科学版, 2000(4):31-34.

[12]朱世武, 崔嵬, 谢邦昌. 移动电话客户流失数据挖掘[J]. 数理统计与管理, 2005, 24(1):62-68.

[13]杨树莲. 数据挖掘在电信行业客户流失分析中的应用[J]. 计算机与现代化, 2005(2):109-111.

[14]C. B. Bhattacharya. When customers are members: Customer retention in paid membership contexts[J]. Journal of the Academy of Marketing Science, 1998, 26(1):31-44.

[15]Wei C P, Chiu I T. Turning telecommunications call details to churn prediction: A data mining approach[J]. Expert Systems with Applications, 2002, 23(2):103-112.

[16]Mozer M C, Wolniewicz R, Grimes D B, et al. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry.[J]. IEEE Transactions on Neural Networks, 2000, 11(3):690-6.

[17]王雷, 陈松林, 顾学道. 客户流失预警模型及其在电信企业的应用[J]. 电信科学, 2006, 22(9):47-51.

[18]Reichheld F F, Jr S W. Zero defections: quality comes to services.[J]. Harvard Business Review, 1990, 68(5):105-11.

[19]夏国恩. 基于核主成分分析特征提取的客户流失预测[J]. 计算机应用, 2008, 28(1):149-151.

[20]邹竞, 谢鲲. C4.5算法在移动通信行业客户流失分析中的应用[J]. 计算技术与自动化, 2009, 28(3):98-101.

[21]陈捷. 基于决策树的电信业客户流失分析[J]. 赤峰学院学报:自然科学版, 2010, 26(12):79-80.

[22]郝梅. 基于CART二叉决策树的电信业客户流失的模型构建与控制[J]. 科技通报, 2012, 28(6):103-105.

[23]邓全. 决策树算法与客户流失分析[J]. 西安邮电大学学报, 2013, 18(3):49-51.

[24]赵冬梅, 刘贵全. 基于时序模式匹配的k-近邻分类在流失预测中的应用[J]. 计算机工程, 2006, 32(10):274-276.

[25]Wei C P, Chiu I T. Turning telecommunications call details to churn prediction: A data mining approach[J]. Expert Systems with Applications, 2002, 23(2):103-112.

[26]朱世武, 崔嵬, 谢邦昌. 移动电话客户流失数据挖掘[J]. 数理统计与管理, 2005, 24(1):62-68.

[27]王维佳, 缪柏其, 魏国省. 数据挖掘——电信客户流失分析预测[J]. 数理统计与管理, 2006, 25(4):419-425.

[28]Rosset S, Neumann E. Integrating Customer Value Considerations into Predictive Modeling[C]// IEEE International Conference on Data Mining. IEEE Computer Society, 2003:283-283.

[29]Yang Y M, Wang H, Li L, et al. Multi-dimensional model-based clustering for user-behavior mining in telecommunications industry[C]// International Conference on Machine Learning and Cybernetics. 2004:1650-1655 vol.3.

[30]尹婷, 马军, 覃锡忠,等. 贝叶斯决策树在客户流失预测中的应用[J]. 计算机工程与应用, 2012:125-128.

[31]朱志勇, 徐长梅, 刘志兵,等. 基于贝叶斯网络的客户流失分析研究[J]. 计算机工程与科学, 2013, 35(3):155-158.

[32] 罗彬, 邵培基, 罗尽尧,等. 基于蚁群算法的成本敏感线性集成多分类器的客户流失研究[J]. 中国管理科学, 2010(3):58-67.

[33]Datta P, Masand B, Mani D R, et al. Automated Cellular Modeling and Prediction on a Large Scale[J]. Artificial Intelligence Review, 2000, 14(6):485-502.

[34]Hwang H, Jung T, Suh E. An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry[J]. Expert Systems with Applications, 2004, 26(2):181-188.

[35]Ferreira J, Vellasco M B R, Pacheco M A C, et al. Data Mining Techniques on the Evaluation of Wireless Churn.[C]// ESANN 2004, 12th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 28-30, 2004, Proceedings. 2004:483--488.

[36]赵红星, 郑丽英, 胡新海. 基于RS\_RBF的电信企业客户流失预测[J]. 首都师范大学学报:自然科学版, 2013, 34(5):43-47.

[37]王学文, 单其帅, 魏彦凤. 基于BP神经网络的电信客户流失风险预测[J]. 科技视界, 2013(28):113-114.

[38]田玲, 邱会中, 郑莉华. 基于神经网络的电信客户流失预测主题建模及实现[J]. 计算机应用, 2007, 27(9):2294-2297.

[39]蒋国瑞, 司学峰. 基于代价敏感SVM的电信客户流失预测研究[J]. 计算机应用研究, 2009, 26(2).

[40]方磊, 马溪骏. 基于信息熵的改进型支持向量机客户流失预测模型应用研究[J]. 情报学报, 2011, 30(6):643-648.

[41]邱一卉, 林成德. 基于随机森林和单类支持向量机的电信行业客户流失预测[J]. 厦门大学学报:自然科学版, 2013, 52(5):603-608.

[42]王纯麟, 何建敏. 基于AdaBoost的电信客户流失预测模型[J]. 价值工程, 2007, 26(2):106-109.

[43]征荆, 丁晓青, 吴佑寿. 基于最小代价的多分类器动态集成[J]. 计算机学报, 1999, 22(2):182-197.

[44]陈晔, 覃晓群. 基于组合预测的电信客户流失预测[J]. 计算机与数字工程, 2015, 43(4):680-683.

[45]孟飞翔. 基于集成分类算法的电信客户流失预测研究[D]. 东南大学, 2007.

[46] Menard S. Applied logistic regression analysis[J]. Technometrics, 1996, 38(2):184-186.

[47] Quinlan J R. Induction of decision trees" Machine Learning[J]. in Data: Goals and General Description of the IN L.EN System." in, 1986:257--264.

[48] Quinlan J R. C4.5: programs for machine learning[M]. Morgan Kaufmann Publishers Inc., 1993.

[49] Lewis R J. An Introduction to Classification and Regression Tree (CART) Analysis[J]. Annual Meeting of the Society for Academic Emergency Medicine, 2000.

[50] B. Scholkopf, C.J.C.Burges, and A.J.Smola. Advances in kernel methods: support vector learning[M]. MIT Press, 1999.

[51]蒋宗礼. 人工神经网络导论[M]. 高等教育出版社, 2001.

[52]陈洁敏, 汤庸, 李建国,等. 个性化推荐算法研究[J]. 华南师范大学学报:自然科学版, 2014(5):8-15.

[53]刘建国, 周涛, 汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展, 2009, 19(1):1-15.

# 致谢

研究生阶段的学习生涯无比美好而且充实，在北京大学软件与微电子学院的学习眨眼而过，无比的留恋这段美好的时光。

本文在论文的选题、写作的不同阶段都受到我的导师李杰教授的悉心指导，当我在写作过程中遇到各种困难时，李杰老师都为我指明了方向，对我精心点拨，把老师丰富的知识传授于我，帮助我开拓研究思路，提升知识水平。同时李杰老师一丝不苟的治学态度，踏踏实实的研究精神，对学生们的关爱引导，不仅让我知道该如何学习，而且教我明理做人，更是让我对老师的胸怀佩服不已，在北京大学软件与微电子学院的求学生涯虽然短暂，却让我感到终身受益，对李老师的感激之情是发自肺腑无法用言语表达。

正是李老师悉心的培养和指导，使我研究问题的思路不断拓展，研究问题的方法愈加科学，对问题有了更深刻的理解，在这样的基础上我的理论水平、知识能力才得以不断提高。

我非常感谢我的父母以及所有的家人，是他们一直在支持我、关心我、鼓励我，陪我度过人生中的困难。他们的理解和支持是我一生的宝贵财富，在此表达我对他们的深深敬意、感激之情。

我还要感谢北京大学智能感知实验室的兄弟姐妹以及所有的朋友们。数载相处中，让我也获得了学校生涯最宝贵的友情。

最后，还要感谢参与本文评审的各位专家学者，感谢他们百忙之中抽出宝贵的时间审阅本文。

# 

# 北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

* 按照学校要求提交学位论文的印刷本和电子版本；
* 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
* 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
* 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名： 导师签名：

日期：