北京大学

硕士研究生学位论文

题目：电信用户离网预测与信用评价

姓　　名：	宋 兵 霄
学　　号：	1201220835
院　　系：	软件与微电子学院
专　　业：	软件工程
研究方向：	电子商务与物流
导师姓名：	李杰 教授

二〇一五年七月

版权声明

摘　要

目前，在电信运营商管理决策中用户信用管理是被用来预防用户消费欠费及恶意消费的最主要手段。用户的信用管理是通过信用评估、分级和授信等手段，来对用户进行分级管理，根据用户信用级别采取有针对性的风险规避及控制措施，来有效降低风险和减少损失。

本文以电信行业中的 A 公司为例，将通过两个阶段对信用评估及授信进行研究分析。第一阶段是对即将离网用户的预测（高危用户），第二阶段是在分离高危用户的基础上构建用户信用评价模型。两个分模型共同构成了动态信用评估的主体。

在用户的离网倾向预测评估中，通过研究对比决策树各种分类算法的不同选择最优算法对某公司的用户进行分类，以分离出高危离网用户。用户信用评价模型在分离高危用户的基础上，对用户的根据消费属性变量进行因子分析得出解释性较强的指标，结合着用户的固定属性变量，来进行两步聚类，对用户进行聚类，分别对每个类别设定不同的信用等级，对每个等级赋予一定的基础信用额度。把每个用户新变更的业务套餐的对信用额度的影响通过风险因子引入其中作为临时调整值。基础信用额度与临时调整值共同构成了一个用户的动态信用额度，每个用户享有的信用额度大小将通过用户历史行为与新套餐的变更共同决定，更加合理可信。

本文的主要贡献包括：（1）动态信用评估首先分离高危用户，对一些即将离网的用户进行预测，不再赋予高危用户信用额度，以减少带来的损失。（2）动态信用评估把用户的新套餐的变更加以考虑，使得用户的信用额度随着业务套餐的变更而变更，形成实时的信用额度。（3）引入风险因子，为用户新办理套餐赋予实时的动态信用额度

关键词：决策树算法、离网预测、两步聚类、信用评价、动态信用评估

The users off-grid prediction and credit evaluation model in Telecom industry

Song Bingxiao(Software Engineer)
Directed by Professor Jie Li

ABSTRACT

Nowadays in the telecom management and decision system user credit management is the main means being used top malicious consumption of a user. Credit management of a user is through credit evaluation, grading and credit and other means, to carry on the classification management of users. Take targeted risk avoidance and control measures according to a user's credit level could effectively reduce the risk and the losses.

The project takes A company in the telecommunications industry as an example in this paper. Credit research and analysis was through two stages. Firstly, predict the off-grid users (high-risk users) among the whole users. Secondly, build the user credit evaluation model based on the separation of risk on users. The two separate models constitute the dynamic credit evaluation model.

The churn tendency prediction model was built through comparison between several different Decision Tree algorithm to select the optimal classification to separate the high-risk off-grid users. User credit evaluation model was built through Twostep clustering algorithm to classify the users by each use's variables which combined both regular variables and factors extract from factor analysis. One could get its line of credit according to which category he is belonging to. The line of credit was named basic credit line. The effect of each new service package to line of credit was introduced by risk factor named temporary adjustment value. A use's dynamic credit line was decide by basic credit line and temporary adjustment value. Each user's credit limit was set through use's historical behavior and the change of new packages.

Main contributions of the paper include: (1) The dynamic credit score model separate the high-risk users first with no grant for those who were predicted to be the off-grid users.(2) The dynamic credit score model take the change of new business package into consideration.(3) Set different variables different weights named risk factors.

KEY WORDS: The decision tree. Off-grid prediction. Twostep clustering. Credit evaluation. Dynamic credit evaluation model

目录

图表目录

第一章 绪论

1.1 项目背景

据工信部数据称，中国的手机用户量目前为止已接近 13 亿人，居全球第一，并且电信行业在国民经济的发展中发挥着相当巨大的作用，电信行业的蓬勃发展推动者国民经济的发展。随着电信业务的发展，手机用户的信用问题越来越突出，相当多的用户在使用电话通讯服务后拒绝支付话费。运营商由于无法追回的欠款而造成了不小的损失，目前各家运营商均有不少的欠款没有追回，同时对消费者来说，由于偶尔的业务套餐变更而未及时缴费而造成欠费停机，给许多忠实客户带来了诸多不便。如在电信行业中企业，一般情况的客户月流失率在 2%-3%左右，如果静态计算，那么所有客户将会在 3~4 年内流失，因此对于运营商来说，哪怕降低 1%就意味着你可以在运营商的市场当中抢占巨大份额，因此，运营商如何通过减少客户的停机次数而挽留住这些用户也就决定了公司业绩。

以 A 公司为例，目前现有的公司信控系统中，对用户信用额度采取的策略是简单生硬的信用控制策略。它只是在每个用户入网初期为用户赋予一个初始入网固定信用额度和动态评估固定信用额度，只是简单的分为三类，分别享有不同的信用额度，对用户没有一个良好的区分，所以信用较好的用户并不会比信用较差的用户多享有更多的服务和优惠，当一个资质较好的用户因为偶然的行为失误而就给用户停机，那么会给用户带来很多的负面影响，为什么我一直没有过违约情况而信用额度只是跟信用较差的用户享有一样的信用额度，影响消费体验，需要进行改进。

现有的信用额度的设置一方面无法对恶意欠费的客户进行分离，任由一些消费者拖欠巨额费用，而没有给予相应的限额，给联通某省分公司带来了较大经济损失。另一方面当用户预存话费使用完，或用户欠费等，月初就会停机，然而对于绝大多数忠实用户，可能是由于一些原因（新产品订购及业务办理，未及时缴费等）造成的停机。目前现有的信控体系无法进行详细的区分用户，给一些忠实的联通用户带来了不必要的损失。如根据联通某省分公司数据可得，平均每月有 400 万次手机的停机，这对联通某省分公司留住忠实客户、提升用户满意度带来了诸多影响，如果联通某省分公司能针对不同用户进行分类，赋予不同的信用等级产生相对性的信用额度，将会极大的改善现在的状况。

本文意在通过分析现有信控模型局限性的基础上，通过分析现有数据挖掘算法在电信行业的适用性的基础上，采用最合适且易于解释的数据挖掘方法，根据现有联通 A 省分公司的用户数据，通过大规模的数据处理构建新的用户预测与信用评价模型，来对上述两种问题寻求解决方案，以来分离出高危用户和根据历史消费情况给予用户

不同的信用额度，以此来减少 A 省分公司的用户停机次数，减少不必要的停开机，帮助联通 A 省分公司留住忠实的用户及增加更多的盈利，提高经济效益。

1.2 国内外研究现状及相关工作

1.2.1 国外研究现状

本文所研究的信用额度在不同行业都有不同的含义，在电信行业的意义就是允许用户欠费的最大限度，当用户的消费行为造成欠费就会给电信的用户停机或者采取其他措施来控制用户继续消费。

随着信用问题被越来越多的行业重视，有关信用评分的理论研究也在飞速发展，研究的重点在于如何把在信用决策中的主观因素变得越来越客观，通过以往的历史数据来对用户的行为进行研究，随着研究的深入，越来越多的统计类算法和非统计类算法被引入到信用评分模型中，随着计算机科学技术的发展和越来越多的数据挖掘新方法的出现，数据挖掘工具如 SAS 企业级，SPSS 等软件，不仅提供的传统的研究方法还有创新的预测模型和诸如决策树、神经网络、支持向量机等不断的被应用到该领域内。

现在的信用评分研究中，已有不少学者将数据挖掘算法如决策树算法引入到该领域，其中有 Raiffa&Schlaifer，D.Sparks， Coffman，Carter and Catlett，Mehta 等[1]，其中较为著名的学者 Makowski 在这方面最早由研究并做出了巨大贡献[2]。Boyle 通过对不同的算法作比较，然后得出了在信用评分方面，分类树有其不可替代的优势。神经网络在这方面的处理有其特有的优势就是可以处理数据结构较为复杂的情况，但是它的结果差别较大，准确度不够高[3]。Yong-Chan Lee 使用支持向量机方法预测公司的信用等级取得了较好的结果。

信用评分模型主要应用在两个领域，一个是金融领域，用于对用户的信用给予一定的评级，对应的用户可以获得一定的借贷款额度及获得某些金融服务等，在过去几年里，国内商业银行通过开展内部风险计量体系建设和实施新《巴塞尔资本协议》，开发了内部评级评分等各类信用风险模型，再者就是电信领域，运营商需要根据用户的历史数据赋予用户不同的等级评级，使得客户在欠费额度以及停机方面分别享有不同的额度，目前的评级主要是凭借总部的评分策略来确定用户的信用等级，而该评分策略没有考虑到不同省份的用户的不同的消费行为和用户属性。

线性判别分析和对数回归模型是在构建信用模型中最常用到的(Abdou, Pointon, & El-Masry, 2008; Desai,Crook, & Overstreet, 1996; Gao, Zhou, Gao, & Shi, 2006; Hand &Henley, 1997; Thomas, 2000; Vojtek & Kocenda, 2006)均做过在这方面的研究[4]。

在客户流失预测方面，Bingquan 利用了七种数据挖掘方法来应用到该方面的研

究中，包括对数回归，线性分类，贝叶斯，决策树，神经网络等，经测试发现均比现有的模型预测准确率要高，在他的模型中 C4.5 和支持向量机的算法效果突出。

Guangli Nie 等把两种数据挖掘算法应用到信用卡用户流失模型中，模型对数据进行了处理，不是把从银行得到的关于用户的 135 个变量完全纳入模型，只是选取了较有意义的变量纳入，同时也把两种算法的误差也考虑进去，建立了一个较为可靠的信用卡流失用户预测模型，在这个模型中对数回归算法与决策树算法被证明在分类方面有着强大的预测能力。

1.2.2 国内研究现状

近些年，由于消费习惯的改变使得越来越多的年轻人开始透支信用卡提前消费，同时，各种信用评分模型开始被广泛运用到信用卡领域，来帮助银行处理一些管理问题，及提前应对一些可预知的信用危机，一种创新型的信用评分模型垂直装袋决策树模型（VBDTM）由厦门大学的张德富老师提出，VBDTM 得到一个聚合的分类器预测属性的组合，通过数据库的测试取得良好精度[5]。

台湾 Ling-jingKao、Chih-ChouChiu 等学者认为进行用户信用打分的关键在于历史消费数据的分析，贝叶斯潜变量模型是通过相当多的限定来产生信用评级的更好规则，相比于判别分析、回归分析、神经网络、多元自适应回归和支持向量机，用贝叶斯潜变量模型在预测客户的类型方面有着较高的准确率，并且影响因素较少，相比于其他几种方法有着较低的错误率。

姚琦云针对电信行业建立的用户信用度等级是从提高运营商欠费催缴率角度出发，并且核算标准重点考虑了缴费情况，而运营商本身的多是采用金融业信用管理办法，不太适合用来作为用户的综合信用评价标准[6]。陈大峰等把模糊数学的方法引入到客户的信用度评定中，给出了对用户评级的评判方法，但是并没有现实的数值验证方法的有效性。

国内的学者王丽平、李多全在结合着美国教授 Satty 的层次分析法的基础上，结合电信行业，通过抽取影响每一个用户信用额度的相关属性，来进行信用度的评估，并把结果应用到对用户的信用控制当中，并通过现实的数据进行实验，验证了模型的有效性[7]。

国内学者在信用评价方面的研究多是侧重于银行有关用户的信用评估。银行业的信用评估与电信行业有关用户的信用额度确定有一定的相似之处。银行业的信用评估是为了减少银行在借贷业务中的风险，对整个经济社会的发展都是有重大影响意义的，相对来说也发展的相对成熟，本文中的电信行业的用户信用额度评估是侧重于运营商的策略目标和用户的消费体验，银行业信用评估的优秀方法可以被借鉴来分析电信行业的信用评估。

1.2.3 研究目标、内容及意义

本文研究目标是研究和建立对 A 公司具有现实意义的动态信用评估，该模型可以在 A 公司现有信用控制模型的基础上能更精确的为用户授予一定的信用额度，为 A 公司做出用户信用评价的模型理论与方法上的指导，并结合已获得的数据进行实证研究。

本文研究内容是通过两个主模型展开，离网倾向预测评估（预测高危用户）与信用评价模型，其中离网倾向预测评估是通过对基于分类算法的不同类型的决策树算法的研究对比，得出关于 A 公司信控系统中分离高危用户最适合的分类器模型，采用 A 公司 3G 后付费用户 2014 年 1 月至 7 月的数据做测试验证模型效果。信用评价模型在分离高危用户的基础上，通过用户的消费属性变量进行因子分析，得出综合因子联合用户的属性变量进行综合聚类。为产生的不同类别赋予不同的信用等级。通过引入风险因子，把用户新办理的业务及临时的通话消费异常对信用的影响加以考虑，得出用户的消费异动额度，综合构成动态信用评估。

数据来源：数据是联通 A 省分公司 3G 后付费用户的数据，总共是 340 万左右。

数据处理：采用 mysql 进行数据的处理，把各种格式的数据通过处理整理成标准格式的可用数据。

模型的构建与测试：在研究各种算法的基础上，在高危用户预测部分通过 Spss Modeler 采用不同的算法对所获取的数据采用 50%做训练，50%做测试来进行模型的构建。在用户信用评估方面采用因子分析和聚类分析。

模型的评估：一方面通过模型预测的准确率来评估模型的优良，一方面通过与现有信用评估模型的对比得出。

本文研究路线图如下：

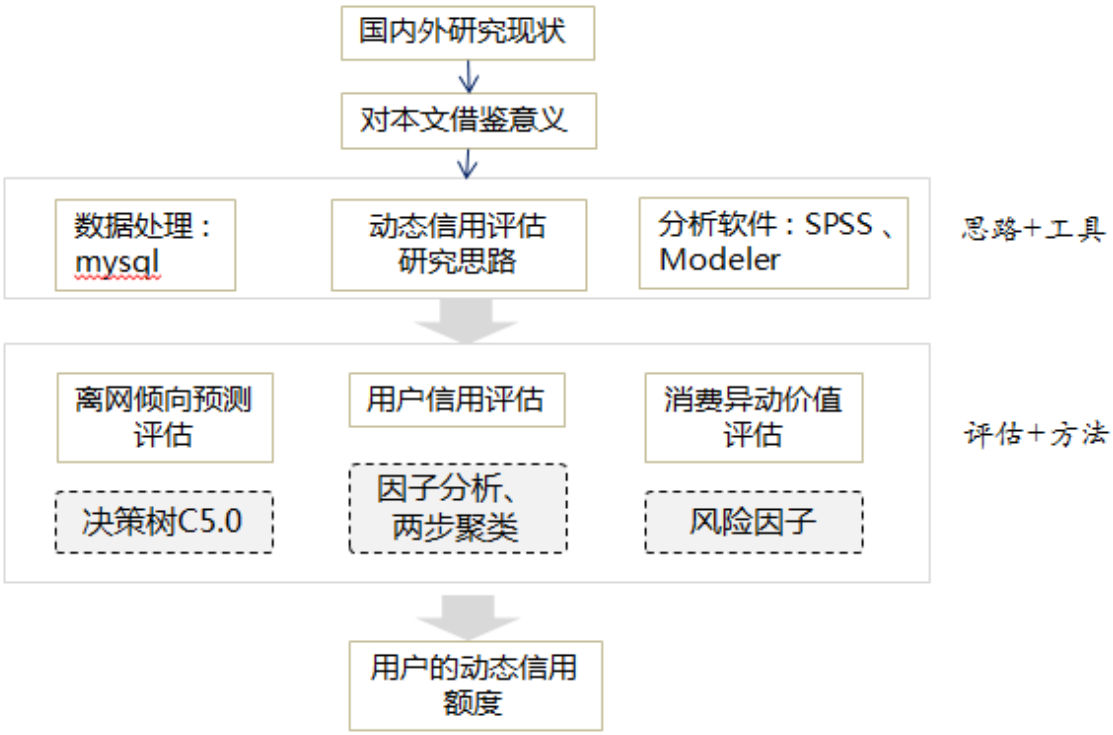

图 1.1　研究路线图

通过对 A 公司数据库中现有数据，包含用户各种属性变量的数据、如话单数据、流量数据 、短信数据、及关键词 、充值时间及渠道 、入网信息明细 、用户场景分类 、已订购产品等属性的数据进行分析，从中提取出关键性数据，通过数据建模构建新的动态停开机预警模型，发现对评估客户信用有价值的属性变量，构建动态的、实时的停开机策略模型，通过把评估出来最优的适合的挖掘算法将这些有用的数据转换成评估客户信用的判断规则，在剥离高危离网用户的基础上，分别每个用户产生的得分聚类，使不同类别享有不同的信用等级，赋予客户实时的信用额度，以减少用户不必要的停开机，及对公司主营业务收入影响，提升联通某省分公司的服务质量，提升用户满意度。

本文的重点在于实时的动态信用评估额度的确定，通过为用户赋予基础额度（用户信用评估得出）和临时调整值（消费异动价值评估）得出每一个用户实时的动态信用额度，使用户享有的信用额度与他历史消费情况和每个月的消费保持一种动态的相关，当用户某月消费额度突然骤增时，会为这样的用户赋予一定范围内的信用额度的授信，以不至于一个用户因为突然的个人消费行为而停机，通过使每个用户享有的信用额度与它的个人消费保持这种保持动态相关，有效的减少用户的停开机次数，增加用户的消费体验，增加运营商的好评度。

1.2.4 本文组织结构

本文共六章，各章内容简要概括如下：

第一章介绍了本文的背景内容，包括目前联通某省分公司现有信控模型的局限性，简单介绍了数据挖掘算法在国内外的信用评价模型中的使用现状，并说明了本文的研究目标设计目的及现实意义等。

第二章分析已有的离网用户预测及其采取的方法和在信用评价方面已有的模型及其比较分析对比，进而引出本文的动态信用评估的研究思想。

第三章主要是对离网倾向预测评估中用到的决策树不同算法的研究，对比得出C5.0 算法是本文中最适合的分类决策树算法，及用 SPSS Modeler 建模的过程及通过 SPSS Modeler 评估方法中的增益图与提升图、提升度和损失矩阵来对对模型的建模效果进行评估。

第四章是对用户信用等级评估及消费异动价值评估的研究，用户信用等级评估方法采用因子分析和两步聚类算法，消费异动价值评估是通过引入风险因子来计算。

第五章是对包括三个子模型的动态信用评估进行总评估。

第六章为总结与展望，总结本文内容并提出相关改进建议。

第二章 动态信用评价理论综述

本章主要介绍项目所涉及的模型研究的理论基础。所涉及到的包括离网用户预测模型的理论基础，在传统行业中被用来预测流失用户的方法。其次包括信用评分发展采用的数据挖掘方法及理论基础，信用评分模型中方法的比较分析，各种模型存在的问题和不足分析，及本文的动态信用模型的基本研究思想。

2.1 离网用户预测理论基础

从通信行业的特点来看，针对流失用户可定义如下：一、主动离网的用户；二、本月没有出账金额并且在 3 个月后可以确定离网的用户（用户三个月没有出账金额才能判断为离网）。本文中的离网用户主要指的是第二种情况下的用户流失，即从当前联通公司转向其它运营商公司[8]。

随着业务的增长，电信市场日渐饱和，运营商之间的竞争更加激烈，用户的选择更多，转网的事件经常会有发生，根据美国市场营销学会的顾客满意手册统计数据显示，获取一个新用户的成本是留住一个老用户成本的 5 倍，运营商开始把更多的目光投向如何留住老用户，“盘活存量”，因此如果能提前预知哪些情况的用户可能会流失，就及早采取措施来留住用户。

客户流失预测是通过离网调研和数据挖掘算法，捕捉客户离网前的特征行为，通过对流失客户特征描述做出预测，其中流失用户特征分析就是利用决策树算法，对流失客户特征进行分析，然后通过这些特征去匹配当前在网用户的特征进而预测该用户是否在下一个月离网。通过对用户的流失分析得到流失用户的数据和潜在可能会流失的客户数据，将这些数据分配给相关的部门，针对流失用户具体特征做出分析，设定个性化的套餐或服务，召回流失客户，挽留忠诚及优质客户。

2.1.1 基于不同方法的离网预测模型

目前已经有不少专家学者在把包括数学方法、统计方法等应用到用户行为分析问题中，通过现有的技术进行建模来预测用户流失情况，较为常用的模型包括回归分析模型、决策树模型、贝叶斯分类模型[9-10]。

一、基于回归分析的预测模型

回归分析是被广泛应用的预测技术，目的是为了找出变量间的依赖关系，并通过函数关系表现出来，回归分析的预测效果仅仅依赖于预测的变量与其他变量的关系，

模型的精确度取决于自变量和因变量的分布符合模型的分布，自变量与因变量的分布适合所选取的模型，则预测效果就好，否则，预测效果较差。Logistic 回归在对商业银行的客户流失方面的预测表明有良好的预测效果，回归分析在客户流失预测中的主要缺点是对训练样本的要求，不能以符号化或者比较直观的易于理解的形式表达隐含的模式。余文建、沈益昌和杜洋用收集到的客户资料拟合了 Logistic 模型，设计了客户基本分、计分标准、最终得分以及信用评级标准，在充分考虑历史信息的基础上做向前预测[11]。

二、基于决策树的预测模型

决策树是比较流行的分类算法，它的构成包括两个阶段，通过训练集生产决策树，然后对生产的决策树进行剪枝，在利用决策树对新样本进行分类时，从树根节点开始对样本进行测试，根据测试结果确定下一个节点，直到到达叶节点，所属类别就是新节点的预测类别。

Kitayama 通过基于决策树的方法对客户信用档案进行分类，把客户群体分等级，分为优质客户和一般客户，接着使用决策树根据客户特征分类，识别高价值客户，以达到挽留高价值客户的目的，决策树在预测模型精度有不错的表现，并且易于解释。

三、基于贝叶斯分类的预测模型

贝叶斯分类是典型的统计学分类方法，被用来预测样本是属于某一特定类别的概率，包括有朴素贝叶斯分类和贝叶斯网络两种方法，朴素贝叶斯基本思想是基于贝叶斯的公式和简化以后的假设，并且根据属性和类别的综合起来的联合概率来估计新样本的所属的类别，类条件独立是应用朴素贝叶斯前提条件，贝叶斯网络是适用于不独立的联合条件的概率分布。

客户流失预测技术的对比研究表明，朴素贝叶斯的预测效果可以和决策树和神经网络相媲美，在电信行业的案例中首先对引起电信客户流失的的因素进行分析，确定先验知识，根据先验知识选取特征和训练样本，通过贝叶斯网络结构学习和参数学习，建立客户流失模型[12]。

2.1.2 基于不同方法的离网预测模型比较分析

在客户流失预测模型中，决策树算法是最常用到的算法，并且在预测精度和可解释性方面都不错，贝叶斯算法由于它是在各个特征属性相互独立的情况下提出来的，这在模型中是很难实现的，因为在电信用户的流失预测模型中，一些变量一定不是独立的，这会导致以贝叶斯算法建立模型会影响预测效果。这些模型都是以静态数据作为训练集，得出模型，再通过另外部分的静态数据来作为测试集，检验模型准确性，通过模型评估来修正模型，满足条件后应用。刘光远、苑森淼等把用户是否投诉作为

一个因素考虑到模型中，通过历史数据与短期的偶发数据，提出链型数据挖掘的方法，联系决策树，形成了链型树分类器，建立了基于该分类器的用户流失预测模型[13]。

2.2 信用评价

信用评分是机构依据一定的公正、科学、权威的资信考核标准，建立对客户信用评估的模型，以此来对客户的诸多方面历史消费行为、个人消费偏好等方面进行综合分析和评价，测定客户履行各种契约的能力和可信任程度，客观的对每个客户的信用进行评分。

信用评级是依据信用评分及运用科学的评级方法，履行既定的评级程序，就企业在未来履行自己承诺的意愿和偿还能力进行判断，赋予每个用户相应的等级，在本文中，主要是对客户历史数据的分析评估。

信用评分最早开始于上个世纪中期，逐渐形成两种体系信用评分：申请人评分和行为评分（Anil Gupta）[14]，信用评分最初是使用统计学方法来区分优质的和不良的贷款，在最初的时候信用评分的重点是是否要给贷款者发放贷款也就是申请人评分（applicant scoring），是应用在金融领域，贷款者会对申请者在未来 12 个月的违约情况进行评估，是利用通过的优秀贷款与不良贷款比例计算得来，以及申请者贷款 1-2 年的数据，相应的信用记录将帮助建立申请者未来两年的申请评分模型。

行为评分是申请人评分的一个补充，是评估申请人在过去一年中支付和购买行为的状况，此数据用于预测未来 12 个月的违约风险，每个月都会更新一次数据，Anil 认为最近表现的借贷信息要比最开始的申请信息更为重要。金融机构的信用评分模型一般会根据每一笔借贷建模，但由于违约情况和借贷款项的不同，到现在为止没有被广泛接受的信用评分模型。

2.2.1 基于不同方法的信用评价模型

信用模型的建立可以采用不同的数据挖掘方法，不管采用哪种方法，均是为了测度用户欠费风险的情况，通过评估用户的属性（包括历史消费、业务套餐等）情况，来为每个用户赋予一定的信用额度，以防止恶意欠费的行为，常见得被用来建立信用评价模型的方法有以下几种。

一、基于人工神经网络的信用评价模型

人工神经网络是基于生物神经元简化的机器学习模型（Hykin）[15]，通过模拟人类的神经元功能，在结构上模拟生物界的神经网络，通过输入层、隐藏层、输出层等来构成，人工神经网络通过模拟生物神经网络在完成一些认知任务时的反应来自行通过计算机改变内置参数来执行，通过对数据进行一系列的计算处理得到结果，通过训

练来得到的非线性预测模型，可以完成诸如分类、聚类、回归分析等数据分析任务。

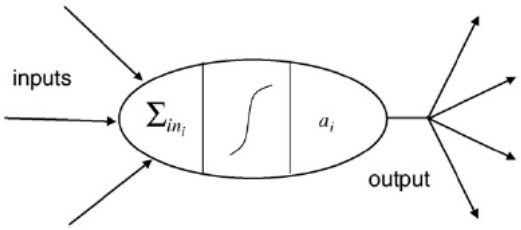

图 2.1　人工神经网络图（1）

Eliana Angelini、Giacomo 和 Andrea 曾在关于神经网络在评估用户的信用方面做了研究，对衡量用户逾期还款的可能性进行建模分析，他首先是在一个小公司的借贷分析中应用神经网络方法来评估一些公司的信用风险，结果被用来作为评级的依据，依据模型而赋予不同的公司不同的信用等级，在他们的研究中阐述了两种基于神经网络的模型，通过广泛的实验分析和依据现实世界的数据来展现了结果。Eliana 等把前馈神经网络应用到了模型中，该模型有一个输入层、两个中间层、一个输出层构成。Eliana 等同时还利用把输入神经元每三个一组连接到下一层，这两种神经网络的应用都是监督学习方法，他应用了这两种方法到信用评估模型中。

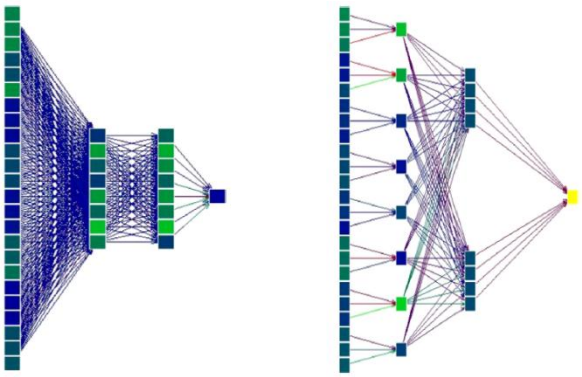

图 2.2 人工神经网络图（2）

人工神经网络在有些领域的应用优于其它的算法，特别是在建模时变量间的关系未知的时候。人工神经网络的算法被广泛应用到经济、金融等众多领域，在金融方面进行信用风险评估方面有着它的优势，同时在银行预测（Tam& Kiang）与股票分类方面（Kryzanowsky, Galler, & Wright）凸显了不足[16-18]。

二、基于线性回归模型的信用评价模型

当中包括一些反应重要信息的随机变量，并且他们被用来当作线性回归模型中的独立变量，不同的变量通过系数（0 或 1）来决定，通过这种方式，他能够识别统计

上有意义的回归系数，模型方程为：

$$Zi=\sum_{j=1}^{n}\beta jxi,j+\varepsilon i$$

其中βj代表 xi 变量重要性程度，这个模型面临的问题是一个借款者违约的概率被假设为从 0 到 1 而不确定何时的概率是多少[19]。

三、基于层次分析法的信用评价模型

层次分析法(Analytic Hierarchy Process)是将与决策总是相关的因素分解成目标、准则、方案等，在这个基础上进行定性与定量分析。它的提出是多目标综合评价方法和应用网络系统理论，是一种层次权重决策分析方法。

1、建立递阶层次结构模型

应用层次分析法 AHP 来解决问题的时候，需要把问题层次化，构造层次结构模型，把复杂的问题分解成元素的组成部分，把这些元素按层次分为三类：最高层、中间层、最低层，这三层相应的代表：目的层、准则层、方案层。上一层元素对下一层元素起支配作用，层次结构中的层数跟所研究的问题的复杂度有关，层次数一般不受限制。每层元素支配的元素一般不超过 9 个元素，因为元素过多会影响两两比较判断的准确度。

2、构造判断矩阵

要比较 n 个因子 $X=\{x_1,x_2,\ldots x_n\}$对某个因素 Z 的影响大小，萨蒂等提出可以对因子进行两两比较形成比较矩阵来判断，每次提取两个因子，用 a_{ij} 来表示 x_i 对 x_j 的影响程度，比较结果用矩阵 $A=(a_{ij})_{m*n}$ 来表示，把 A 称作 Z-X 之间的判断矩阵，从矩阵中可看出，如果 xi 对 xj 的的影响之比被记作 a_{ij}，那么 x_j 对 x_i 的影响就是 $a_{ji}=1/a_{ij}$，萨蒂建议用 1~9 及其倒数来表示两个因子相互间的重要程度

在此步骤上还要通过层次单排序一致性检验和层次总排序一致性检验[20]。

层次分析法具有很强的个人主观因素，得出的结果可能会因人而异，因而不具有代表性。

四、基于 K 均值聚类的信用评价模型

左子叶、朱扬勇把数据挖掘的聚类算法应用到银行的信用卡评级系统 DMCA(Data Mining for Credit card Analysis)，数据挖掘的聚类算法是将样本或者变量分为多个类，把相似度高的类归为一类，把相似度低的样本放在相异的类，是一种无监督学习算法，通过把大数据分类，很好的解决了在信用评分评级中边界值的确定、交叉验证等问题。DMCA 系统中便是采用了 k-means 聚类算法来对采样数据进行聚类，通过分类来得出各个级别的分界值，再通过分界值对所有数据进行分级评定[21]。

五、基于线性判别分析的信用评价模型

线性判别分析模型的经典例子有 Altman 的 Z-score 模型，通过一个判别方程进行多元分析，允许在一个单独的 Z 值中有多个变量的值，和限定值作比较来给予贷款者做判定，允许贷款或者不允许贷款，线性判别分析的模型方程是：

$$Z=\lambda_1 X_1+\lambda_2 X_2+\lambda_3 X_3+\ldots\lambda_n X_n$$

其中 x 代表独立变量， λ 代表判别系数。

此外还有对数回归模型在信用评价中的应用，对数回归模型由于它的鲁棒性和透明性因此在银行业是最为常用的模型，尽管有许多新的技巧在信用评分和在预测精度方面有优越性，但是他们在结果的可解释性上存在困难，因此一些先进的方法无法被广泛应用在现实中，为了提升对数回归的预测准确性，对数回归的随机相关系数被提出了，预计在预测精度方面有不错的效果。

2.2.2 基于不同方法的信用评价模型比较分析

现有的信用评价模型被应用最广的是在金融领域，用户的信用评分被用来评估用户的信用情况，许多不同的方法被用于信用评价模型，不管什么方法被采用，信用评价模型都要解决的一个问题是在一段特定的时间内申请者的违约情况的可能性如何。在上述的几种被用到的方法中，各自有各自的优势，人工神经网络在很多情况下可以模拟现实的世界，它在建模的变量关系未知时效果更好，神经网络可以进行超大量数据的处理，分类的准确程度较高，人工神经网络采用的并行分布处理能力及分布存储能力都很强，对有噪声的数据有很强的容错能力，能够充分的逼近复杂的非线性关系模型的拟合，但是它的不足在于人工神经网络的参数的设置非常复杂，如网络拓扑的权值、阈值等的初始值的设定等。如果模型可以建立，不容易对它的建模过程进行学习，以此得到的结果不容易解释，因此会影响结果的可接受程度，尤其是被应用于解决现实问题时，会引起对结果的质疑。人工神经网络被应用来建模解释现实世界场景的情况不多。

线性回归是比较重要的统计推断方法，在社会发展的很多方面都用到了线性回归模型，线性回归模型是在分析多因素模型时，能准确的计量因素的相关关系和衡量回归拟合程度的准确性。现实情况中，模型变量的重要程度的确定相对困难，选取某种或者某些因子来对因变量做预测很多时候是通过主观因素决定，这就影响了回归模型的不可预测性，使得对于有较多影响因素的回归模型造成很大误差[22]。

层次分析法，是采用系统化的思想，不割断模型中每个因素对结果的影响，其中每个层次中的每个因素对结果的重要程度都可以量化，每一层的权重设置都可以直接或者间接影响到结果，这种方式非常清晰、明确。层次分析法模拟人脑的决策方式，对现实世界的问题采用定量与定性结合的方式，层次分析法需要构造判断矩阵，采用定性成分多于定量的数据来对问题做出判断，因此当大家针对现实问题展开探讨时，

人们经常会对相同的问题持有不同看法，对问题的重要程度的判定个各不相同。因此构造出两两因素相互比较打分判断矩阵的很依赖个人主观性，得出的结果可能不够客观因而不常被采纳[23]。

2.3 动态信用评估的基本研究思想

数据挖掘算法已经被广泛的应用在客户流失预测模型与信用评分模型中，并且已有众多的专家做过这两方面的研究。在客户流失预测模型中分别有采用回归、决策树、贝叶斯分类等方法用来建模。本文的客户流失预测模型通过比较不同分类方法的优缺点后及适用的情况，针对本文的研究目标决定采用决策树分类算法来进行建模，决策树算法又包含各种不同的分类标准属性变量值确定的方法，通过各种算法的优劣对比分析出最合适本文的决策树算法，决定采用 C5.0 来对高危离网用户进行预测。通过对基于分类算法的不同类型的决策树算法的研究对比，得出关于 A 公司信用模型中分离高危用户最适合的分类器模型。本文中的信用评价模型将采用 A 公司内部打分规则计算得到每个用户得分，对所有用户根据得分采用 k 均值聚类，赋予不同的类别不同的信用等级，不同的信用等级可以享有一定的基础信用额度。通过引入风险因子，对用户新办理的业务进行消费异动判断，这样可使得消费高的用户享有更高的信用额度，不会因为新办理业务套餐而停机，赋予用户实时的信用额度，通过临时套餐的变更产生临时调整值来构成实时的消费异动价值。

本文将采用客户流失模型与信用评分模型相结合的方式来构建动态信用评估。

把客户流失预测作为第一步来构造模型，通过预先设定那些用户的行为为高危，对获得的数据进行建模，目的是为了分离高危用户，不再对高危用户赋予信用额度，以避免有些用户的恶意消费而造成公司的损失。

用户信用等级评估是根据每个用户的历史消费行为来对用户进行聚类，分别授予一定的信用等级。每个信用等级享有一定的信用额度。

消费异动价值评估是每个用户的个性化评估，会根据用户实时的通话、上网等各种属性来给用户增加实时的信用额度。

逻辑图如图 2.3：

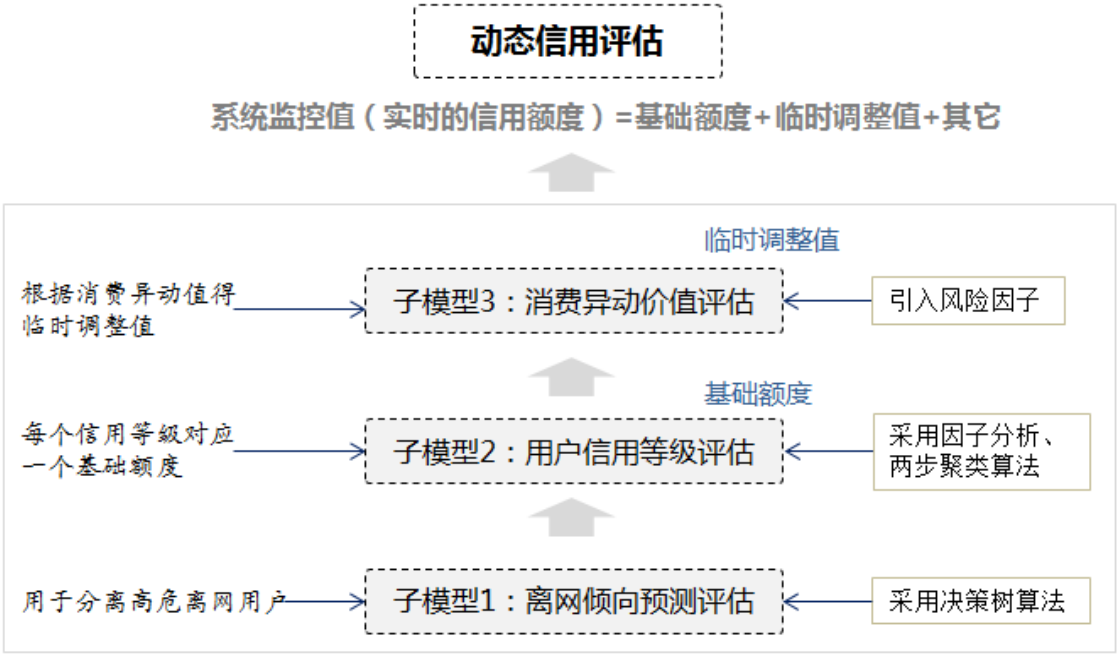

图 2.3　动态信用评估逻辑图

2.4　本章小结

本章第一部分总结了在客户流失方面各领域有过的研究及应用的方法，对比得出决策树在对用户分类方面的优势，在预测精度方面要比贝叶斯网络和人工神经网络要好，并且易于解释。本文的离网预测模型中便采用的是决策树算法来分离高危离网用户。本章后半部分介绍了信用评分和信用评级采用的模型方法，人工神经网络的应用在信用风险评估方面有优势，但是在银行信用评分方面凸显不足，回归分析中的权重的设定没有较为科学的方法，层次分析法是在信用评级方面是不错的选择，因为它是把每层的因素对上层因素的重要程度给予打分，这样使得较为重要的属性享有较高的权重，不足之处在于构造打分矩阵有很强的主观性，有些情况下不能很好的代表事实。k 均值聚类能很好解决分类的边界值确定等问题，本文便是利用某公司内部现有的打分规则，对用户评分进行 k 均值聚类，分别赋予不同的类别不同的信用等级[24]。

第三章 离网倾向预测评估

离网倾向预测是动态信用评估的第一个步骤如图 3.1。

离网倾向预测评估是建立后两个模型的基础，离网即为流失客户中的从当前运营商转网到其它公司的这部分用户，从整体用户中分离出高危的即将或已经离网的用户，从本质上看，是属于分类问题，把所有的用户分为高危离网用户与正常用户。本文是针对电信用户来具体分类，由于可以获得 A 公司关于客户的各种属性数据，并且决策树便是通过判别函数对用户的属性变量值进行分类，对不同的用户类别做了很好的区分，因此离网倾向预测评估便是采用的决策树分类方法来建模。

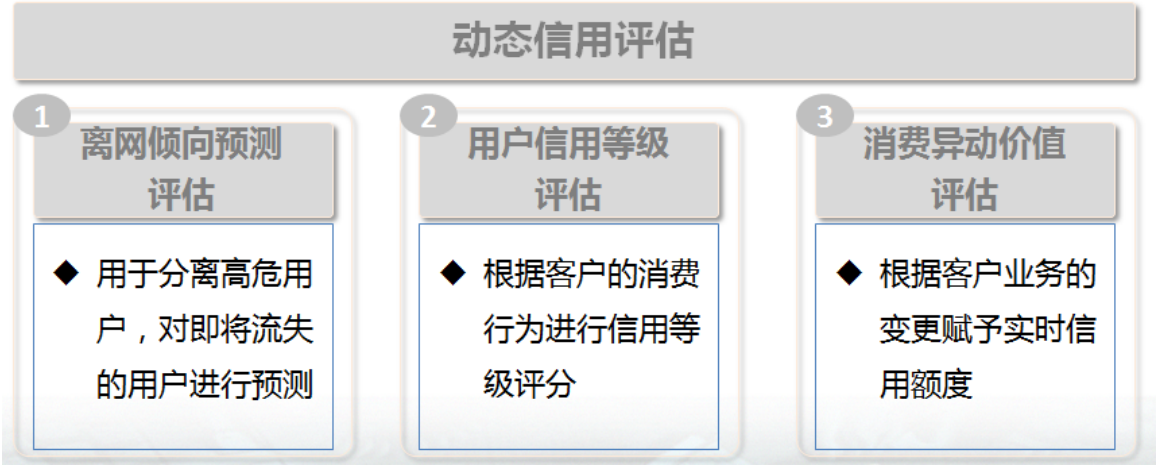

图 3.1 动态信用评估框架图

3.1 模型设计思路及研究方法

构建模型的数据来源于 A 公司 3G 后付费用户的消费数据，从中提取用户的部分属性，对预测高危用户相关的属性，主要为用户的基本信息、消费信息、行为信息等相关的。正常用户定义为：数据统计日状态正常，统计中最近一周有通话记录，当月有出账记录，第三个月月底出账正常。高危离网用户定义：数据统计日为正常的用户并且没有停机，有最近一周的通话行为，截至下月月底连续一周以上无通话行为，下月月底不出帐。

提取的有关用户的属性变量数据如下所示（表 3.1）：

表 3.1　用户属性变量表

用户属性变量	说明
USER_NO	用户手机号码
地市	A 公司所在省份不同城市
是否集团	1 是、2 否
VIP 客户等级	金、银、非 VIP
七大渠道	集团直销、电子渠道、自有实体渠道、社会渠道、社区直销、其他等.
在网时长：	单位：月
合约类型：	普通单卡、存费送机.
是否沃家庭：	1 是、0 否
前六个月平均累计欠费金额：	单位为元
前六个月平均预存款总额：	单位为元
前六个月平均出账金额：	单位为元
前六个月平均同网通话次数：	网内通话次数
前六个月平均本地主叫计费时长：	单位为分钟.
前六个月平均本地被叫计费时长：	单位为分钟.
前六个月平均上网流量：	单位：M
上月累计欠费金额离均差：	用户欠费情况与所有用户平均欠费额度比较值.
上月预存款总额离均差：	单位为元
上月出账金额离均差：	单位为元
上月同网通话次数离均差：	单位为次.
上月本地主叫计费时长离均差：	单位为分钟
上月本地被叫计费时长离均差：	单位为分钟
上月上网流量离均差：	单位为 M.
上月半停次数离均差：	单位为次
上月停机次数离均差：	单位为次
前三个月平均半停次数：	单位为次
前三个月平均停机次数：	单位为次
7 月份是否流失：	1 是、0 否
半年平均异网通话次数：	单位为次
半年平均通话次数：	单位为次
6 月份异网通话次数离均差：	单位为次

表 3.1 字段备注：由于建模阶段采用 1 到 6 月份和 7 月份的流失用户的数据建立模型，因此此模型的作用是通过前六个月的数据预测用户在七月份是否流失，由于 1 到 3 月份用户的停机数据缺失，所以选取 4 到 6 月份的用户停机数据，在模型部署后再对模型优化时，需要提取前 6 个月的数据，由于是求均值，因此缺失数据对模型的

影响不大。建模阶段的数据选取排除了样本数据中 1 到 6 月份为 3 无用户（无通话、无流量、无出账金额）的记录并且排除了 6 个月平均出账金额小于等于 0 的记录，此类用户均归为高危离网用户群。

以上为建模的用户属性变量字段选择，是通过了将大量对建模无关的属性进行了删除，如用户身份证号码、姓名等。只选取了对离网预测模型建立相关性较大的属性变量出来，并对数据做预处理。

3.2 构造决策树方法

3.2.1 决策树

决策树是一种简单并且广泛使用的分类器,它提供一种在什么样的条件情况下就会得到什么样的值分类的规则方法。通过对数据进行处理，利用训练集数据、归纳算法生成可读的规则集和决策树，然后使用决策树对新数据进行分析，实际上是通过一系列规则对数据进行分类的过程。决策树的优点有两大主要方面：一是可读性好，可解释性强，具有描述性，有利于人工分析。二是效率高，只需要构建一次决策树，就可以反复使用，每次预测的最大计算次数不超过树的深度。

决策树是由决策点、分枝和叶子组成的，其中每个内部节点表示在一个单一变量上的测试,这种测试能将整个数据集合分割成两块或者更多分枝,每个分支代表一类，每个叶节点代表都是属于单一的类别记录，树的最顶层节点是根节点，根节点代表的是整个数据集合空间，沿着决策树从上到下不断进行遍历的过程中，在每一个节点处都会遇到一个测试，每个节点上的测试导致不同的分枝，最后到达叶子节点。

决策树可以分为两种：分类树和回归树。分类树与回归树是对离散变量和连续变量分别做决策树。

决策树的生成主要包括两部分，树的生成和树的修剪。

构造决策树要寻找初始分裂，构造的决策树的集合是通过训练集产生，训练集当中每个案例须是已分好类的，决定如何把哪一个属性域当做作为最好的分类，一般是通过穷尽训练集包含所有的属性域来判断，分别依次对每个属性展开的分裂的好坏做出量化，计算得到一个最好的分类，量化的标准是很多样化的，不断重复，直到每一个叶节点的记录都属于一个类中，增长称为一棵完整的决策树的过程是（如图 3.2）：

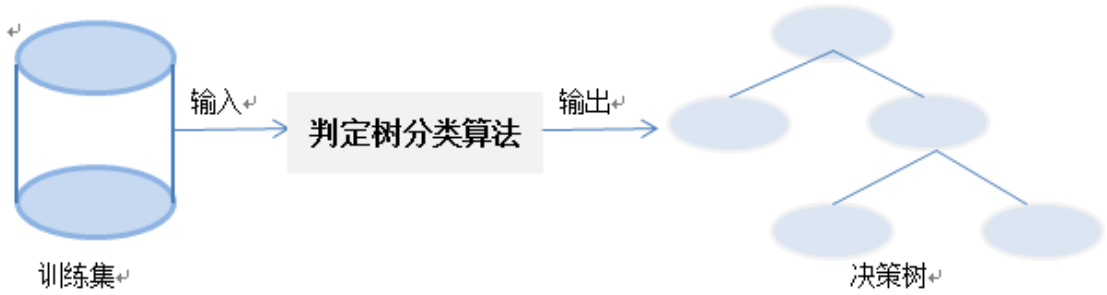

图 3.2　决策树流程图

3.2.2 各算法优缺点比较

ID3 算法存在的不足：ID3 选择属性的原则，通常是选 E(A)最小，这样存在的弊端就是算法往往会偏向选择取值多的属性，而取值多的属性不意味着是最优属性，按照熵最小原则被 ID3 算法列在首要判断条件的属性在现实中有时不那么重要。在分析实际问题中的很多算法是在 ID3 算法基础上进行优化改进。

C5.0 决策树的原理是基于产生最大信息增益的字段来逐级分割样本，优点是在执行效率和内存使用上都有明显改进，非常适合于大数据集，对数据遗漏的问题很稳健，不需要很长的训练次数来进行估计，目标字段必须是分类字段，C5.0 模型比一些模型更容易被理解，模型推出的规则可以很直观的给出解释，允许多次的多子组分割。离网倾向预测评估就是采用的 C5.0 算法来分离高危用户。

CART 基于树的分类和预测的算法，模型简单，较容易理解，规则解释起来更明了，CART 是通过在每一步最大限度降低不纯洁度，用递归分区将训练集分组，根据使用的建模方法自动选择最合适的预测变量，目标变量和预测变量既可以是范围字段，也可以是分类字段，所有的分割均是二元分割，通过基尼系数来判别，但是 CART 算法的稳定性较差。

CHAID 决策树是通过卡方统计量识别最优分割来构建决策树的方法，它可以产生多分支的决策树，目标和预测变量既可以是范围字段也可以是分类字段，是从统计显著性的角度来确定分支变量和分割值，优化树的分支过程，CHAID 是依据目标变量来对输入变量进行多水平划分。

QUEST 决策树运算过程较为简单，QUEST 节点提供构建决策树的二元分类法，相比于 CART 算法更省时，也减小分类树方法中常见到的偏向于类别较多的变量的趋势。对于 QUEST 预测变量字段可以使数字范围的，目标字段必须是分类的，所有分割都是二元的。

通过以上各种决策树算法的对比，及 A 公司现有用户数据的属性变量，发现 C5.0 有较为合适的算法，C5.0 算法具有较强的解释性且易被理解，被用来在子模型一中

分离高危离网用户。

3.2.3 C5.0 算法

C4.5 是在 ID3 的基础上，通过计算信息增益比来选择决策树结点分类属性，在一定程度上克服了 ID3 偏向多值属性方面的不足，还增加了剪枝等，C5.0 是在 Quinlan 通过 C4.5 的基础上改进而得到的新算法，C5.0 算法的执行效率和内存使用均较 C4.5 有改进，一般不需要很长的训练次数，当出现数据遗漏和数据字段问题时表现非常稳健，C5.0 的模型更易理解，允许多次多组的分割，这种算法对于商业大数据非常适合。

决策树的属性选择标准有多种，如有信息增益率，基尼指数与距离度量等，C5.0 采用的是用信息增益率来作为选择属性的标准。

步骤如下：

假设 S 是数据集，类别有{C_1，C_2，…，C_k}，选择属性 R 把数据集分为多个不同的子集。

假设属性 R 有 n 个互不重合的取值{r_1，r_2，…r_n}，那么 S 被分成 n 个子集 S_1，S_2，…，S_n，那么 S_i 中的所有记录的取值均为 r_i。

令|S|代表数据集的数据个数，$|S_i|$代表属性 $r=r_i$ 的案例数，$|C_j| = freq(C_j,T)$,是 C_j 类的案例个数，$|C_jr|$是属性 $R=r_i$ 例子中具有 C_j 类别的例子个数。

有：1) 类别 C_j 的发生概率：$P(C_j) =|C_j|/|S|=freq(C_j,S)/|S|$

2) 属性 $R=r_i$ 的发生概率为：$P(r_i) = |S_i|/|S|$

3) 属性 $R=r_i$ 的例子中，具有类别 C_j 的条件概率：$P(Cj/ri)=|C_jr|/|S_i|$

4) 每个类别的信息熵为：

$$H(C) = -\sum_{j=1}^{k} P(Cj)\log2(P(Cj)) = -\sum_{j=1}^{k} \frac{freq(Cj,S)}{|S|} * \log_2(\frac{freq(Cj,S)}{|S|}) = info(S)$$

5) 类别的条件熵为：

按照属性 R 把集合 S 分割以后的条件熵为：

$$H(C|R) = -\sum_{i=1}^{n} P(Ri) \sum_{j=1}^{k} P(Cj|ri)\log2\, P(Cj|ri) = -\sum_{i=1}^{n} \frac{|Si|}{|S|} * info(Si) = infor(S)$$

6) 信息增益 Gain

$$I(C,R) = H(C) - H(C|R) = into(S) - infor(S) = gain(R)$$

7) 属性 R 信息熵

$$H(R) = -\sum_{i=1}^{n} P(ri)\log2(P(ri)) = -\sum_{i=1}^{n} \frac{|Si|}{|S|} * \log_2(\frac{|Si|}{|S|}) = split\_info(R)$$

8) 信息增益率为：

gain_ratio = (C,R)/H(R) = gain(R)/split_info(R)

属性选择和样本分区大的标准就是最大信息增益率，经结果来看它是比信息增益法更好，克服了其它算法会偏向于取值较多的属性变量。

3.2.4 C5.0 决策树的修剪算法

C5.0采用的是后剪枝方法，从最终端的叶节点依次向上逐层的开始修剪，C5.0采用的剪枝方法中主要是关于误差的估计和如何设置修剪标准。

误差的估算

通常是根据决策树在样本集上的预测误差情况来修剪，C5.0是应用了统计学的置信区间的方法，步骤如下：

1、针对每个决策树的节点，把输出变量的众数类别作为预测类别。

2、设在第 i 个节点包含 Ni 个观测，有 Ei 个是被预测错误的观测，那么误差就是

$$f_i=E_i/N_i$$

3、对第i个节点的真实误差e_i在近似正态分布的假设上进行区间估计，置信度为1-a：

$$P\{\frac{f_i-e_i}{\sqrt{\frac{f_i(1-f_i)}{N_i}}} < |Z\frac{a}{2}\} = 1- a$$

其中$Z\frac{a}{2}$为临界值，C5.0默认置信度是75%，当a=0.25时，$Z\frac{a}{2}$=1.15

置信度越大，则所允许的悲观误差估计越高，剪去的树枝越大，置信度越小，所允许的悲观误差估计越低，被剪去的树枝越小。

修剪标准

在误差估计基础上，C5.0是根据“减少-误差”法来判断是否修剪。

首先计算待剪子树中叶节点的加权误差，与父节点的误差相比较，如果大于父节点的误差则就剪掉，否则不能剪掉

$$\sum_{i=1}^{k} pi ei > e \ ,l = 1,2,\ldots k$$

公式中 k 是待剪子树中叶节点的个数，P_i 是第 i 个叶节点的样本量占整个子树样本量的比例，e_i 为误差估计，e 是父节点的估计误差[25-30]。

3.3 基于 SPSS-Modeler 软件分析

SPSS Modeler 是一组数据挖掘的工具，通过 SPSS Modeler 工具可以针对商业案例建立快速预测性模型，应用到商业活动中，来改变决策的过程，SPSS Modeler 提供了各种利用机器学习、人工智能以及统计学来建模的方法。它能够帮助用户发现隐

含在数据中的信息，提供了利用算法把数据及统计量间关系可视化的窗口，对数据执行的操作可用节点表示，将各个节点连接即可形成一个流，流表示对数据的操作过程，通过图形化的操作使人一目了然对数据操作过程，简单易用。

提取 A 公司系统中的数据，对数据处理的过程数据流（如图 3.3）：

图 3.3　C5.0 算法建模数据流图

源：数据需要通过源节点导入到 SPSS Modeler 的数据流中，本文中是把 SPSS Statistics 数据文件导入到 SPSS Modeler 中。

字段选择：可以对数据字段进行操作，比如进行过滤、数据字段类型的设置、数据的导出、数据的 RFM 分析等。

本文将数据导入 SPSS Modeler 中后，对数据字段进行过滤，在所有字段中共有 33 个变量，其中经过过滤操作后，过滤了 4 个变量，剩下 29 个变量留下进行建模的分析使用。

通过字段选择中的类型设置来对所有字段变量各种属性进行设置，本文中将 29 个变量中的地市、是否集团、VIP 客户等级、七大渠道、合约类型、是否沃家庭、是否流失变量等均设为名义变量，其它变量为连续型变量。把 USER_NO 的角色设定为记录 ID（无实际意义），把最后一项的是否流失角色设定为目标变量。

分区：是为了把所有的数据进行分割，对高危用户的预测需要利用训练集进行训练，用测试集对训练出的结果进行模型的验证，以此验证模型的预测结果是否达到标准，达标后才可在新数据集上进行预测，因此在建模前，要把所有的数据进行拆分，拆分为训练集与测试集。本文中的模型是把所有数据的 50%用来做训练，把剩下的 50%用来做测试（如表 3.2）。

表 3.2　测试集与训练集比例表

训练集	50%
测试集	50%

所选定的训练集应该要尽量完整的包含高危离网用户的属性特征，这样得出的离网预测倾向模型才能完整的体现有离网倾向的用户的特征，因此，需要在训练集中详细的设定正常用户与高危用户的比例。根据 A 公司的现实数据及经验显示，A 公司的客户流失比在 2%左右，这样离网用户与正常用户的比在 1:49 左右，如果把从 A 公司得到的数据直接进行建模则会导致模型训练的不足，对高危用户不能很好的预测，模型会预测所有的高危用户为正常用户，这样建模得到的预测结果准确率貌似很高，但是对建模就没有了实际的意义。建模真正关心的是预测为高危用户的人群中是真正的高危用户的比例，因此需要平衡训练集中的高危用户与正常用户的比例，来使得模型对高危用户的预测准确率更高。测试集不需要进行样本比例的重新平衡，保持原始数据就可，这样通过了训练集建模在测试的过程中，还可以发现模型是否过拟合，从而进行调整。

平衡：是为了调整模型中数据的比例，将高危用户与正常用户的比例进行了调整。

表 3.3　平衡因子表

因子	条件
1	正常用户
1.5	高危用户

在上表 3.3 中可以对平衡节点进行设置，每个平衡指令包含着平衡因子和平衡条件，通过因子和条件来指定因子值比例记录，在本文中是把满足高危用户及有离网倾向的用户条件的记录因子设为 1.5，把正常用户的因子设为 0.1，使两者的比例保持在恰当的范围内，来防止最后建模得到结果的不准确。

建模：构建模型采用的决策树 C5.0 数据挖掘算法。模型的建立采用的输出类型采用规则集，生成推理规则集采用的是 PRISM 算法，不断缩小样本量来获得推理规则。

PRISM是一种覆盖算法，由它生成的规则在训练样本集上是100%正确的。步骤是：

1）在全部观测内，寻找这样一条规则它能最大范围覆盖属于该类别样本推理规则。如果有M个观测，其中N个属于期望类别，是使正确覆盖率（N/M）最大的标准即为确定规则的标准

2）在样本量为M的样本范围内，确定附加条件的最大原则是正确覆盖率最大，得到小一些的样本范围，在此基础上不断附加条件，不断缩小样本范围，一直到推理规则不再覆盖其它类别的样本，便形成了一条推理规则。

PRISM算法就是不断通过逐步缩小样本空间范围得到最终期望类别的样本，同时得到相应的推理规则。

规则集也需要精简，精简的过程是基于测试样本集，针对每条推理规则，找出它所覆盖的所有观测，先去掉一个条件，计算正确覆盖率和误差率，如果误差率低于原规则的误差率则去掉相应的条件，重复继续直到剔除后误差率高于之前为止。

生成推理规则集有两种方法，由决策树直接生成和由PRISM算法生成，结果是一致的，均来自于决策树，由决策树生成的规则集能够覆盖所有样本。

在模型设置中的模式选择上采用专家模型，修剪严重性是决策树修剪时的按置信度为75%来计算，每个子分支的最小记录数是指定每个节点允许的最小样本量为50，当子分支的记录数小于 50 时会被剪掉。

当我们使用SPSS Modeler对商业问题进行分析探讨时，做出的决策是要背负很大商业期望的。如果预测失败，就要承担相当的损失。虽然数据挖掘中误差是不可避免的，但如果把可能出现的误差及其导致的损失反映出来，损失矩阵就为这样的问题提供了很好的解决方法，损失矩阵可以把可能导致的损失引入到模型构建分析过程中，从而得出更加符合实际的结果。

有的时候分类模型给出的分类预测结果可能是错误的，不同类型的错误造成的损失是不同的。二分类模型中，判断错误包括两种情况：一种是实际为真却预测为假，即为弃真错误，带来的损失为m，另一种是实际假却预测为真，称取伪错误，带来的损失为n，由这两类错误的损失矩阵。判断存在失误时带来的损失是不同，判断失误时的成本代价是不同的，如表3.4.

表 3.4　损失矩阵表

		预测值	
		Yes	No
实际值	Yes	0	m
	No	n	0

在两种情况下使用损失矩阵: （1）数据建模阶段；（2）样本预测时使用损失矩阵。在SPSS Modeler的C5.0采用的是在数据建模阶段使用损失矩阵，并不会影响决策树的生长，是在修剪中考虑了损失矩阵，C5.0将按照“减少-损失”法即为判断待剪子树中叶节点加权损失与父节点损失是否更大，当大于父节点损失时时则可以剪掉：

$\sum_{i=1}^{k} pi ei$ ci>ec , i =1,2,…k

其中，k是子树中叶节点的总个数，P_i是第i个叶节点样本个数占整个子树样本个数的比例；ei为第i个叶节点的估计误差，ci规定为为第i个叶节点的错判损失，e为父节点的估计误差，c为父节点错判损失。

本文关于损失矩阵的设置中，把实际为高危用户预测为正常用户的损失设定为2，把实际为正常用户预测为高危用户的损失设定为1，相比较而言，对A公司来说，把高危预测为正常用户带来的损失要比把正常用户预测为高危用户代价大、带来的损失也大，因此设定的损失值也较大。

由于定义了损失矩阵，使得决策树的修剪会更加慎重，要同时考虑误差大小和带来的损失多少，当修剪造成损失较大时，则放弃修剪。

3.4 本章小结

本章是对离网倾向预测评估建模过程用到的算法及建模过程进行了详细的描述，对离网用户的预测本身就是一个分类问题，本文选择采用决策树分类算法应用到模型中，因为决策树算法规则明确，易于解释，并且建模的过程及结果都很容易解释，此模型最终是根据市场部的需求来建立的，所以选择效果好并且容易被理解和接受的算法应用到模型中。

本文对决策树不同算法的优劣情况进行了对比，每个算法均有自己不同的特点和适用的情况，通过对比发现决策树算法中的C5.0算法对离网倾向模型的构建非常合适，因为本文中采用决策树C5.0算法构建第一步的模型，数据来源于A公司2014年的1到6月份的数据，通过对数据的清洗处理，删除无关的属性变量，得到最终被用到模型中的30个属性变量的标准化格式数据，采用SPSS Modeler对标准化数据进行建模，按步骤的经过字段过滤、属性变量角色设定、训练集与测试集的分区设定、平衡因子设定、模型的剪枝标准、损失矩阵的设定，得到了通过离网倾向预测评估对高危离网用户进行判断的规则集，设定“0”为流失用户，通过规则集被判别为“0”属性的用户，即为高危离网用户（流失用户）。

建立的高危离网倾向预测评估要经过模型的评估来确定模型的优良准确性，该模型采用最常用的增益评估图与提升评估图来对模型的建模效果进行评估，该模型的准确率已达到95%以上，可以很好的预测即将离网的用户人群，并且第二个衡量指标是把高危用户预测为正常用户的数量要尽量的小，因为相对于运营商来说，把高危用户预测为正常用户而为用户授信带来的损失要远远大于把正常用户预测为高危用户的实际损失。所以运营商宁可选择牺牲把正常用户预测为高危用户，也要将把高危用户预测为正常用户的精度提高。通过C5.0算法计算的结果中，的重合矩阵可得出把流失用户误判为正常用户的数量是2234，量已经非常小了，通过了模型的评估以确保模型

可以在实际应用中可以使用，将具有良好的现实意义。

第四章 用户信用等级评估与消费异动价值评估

用户信用等级评估是动态信用评估的第二部分。

建立用户信用等级评估是为了对众多的用户根据每个用户的属性变量综合情况进行分类，用于判断用户的信用等级，并且为相应的信用等级赋予相应的信用额度。进行用户的信用评估的建模方法有不少，本文的特点在于把用户的信用等级模型与离网倾向预测评估进行结合，以往的模型是对公司的所有用户根据其属性变量就开始建模，而本文的特点即为先进行离网倾向预测评估，把公司的所有用户进行分类，这样可以根据规则判别用户的分类，把预测为高危离网特点的用户进行分离，在分离高危用户之后，开始对用户的信用评级，这里采用的是对用户的属性变量进行因子分析，从中提取出解释性较强的综合性指标，通过每个用户的综合指标进行聚类，聚出不同的类别，通过聚出的类别，根据实际情况为每个类别赋予一定的信用额度。

本章结构路线图

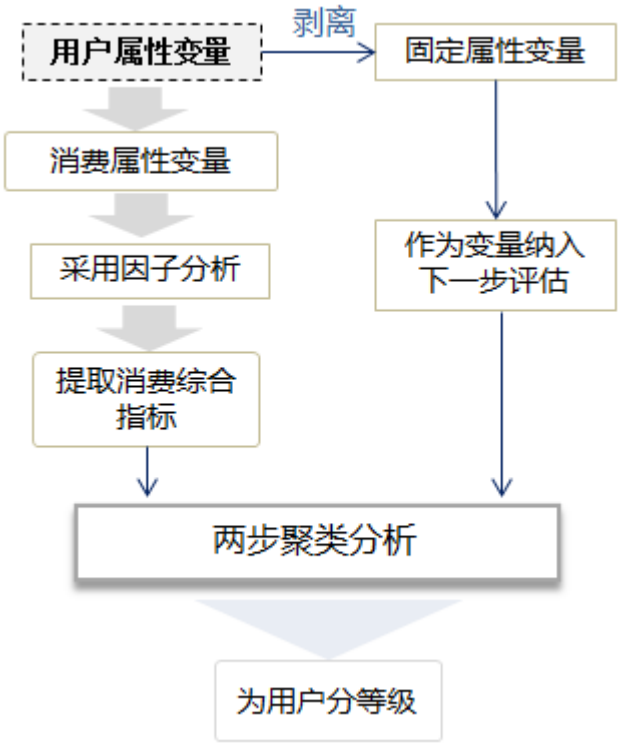

图 4.1　用户信用评估路线图

4.1 用户信用评估相关方法介绍

本文在建立离网倾向模型的基础上，对高危用户分离，对正常的用户采用聚类算法，来确定相似记录的用户并把相似类型客户聚合为一类，并添加类别标签，聚类算法不需要事先了解类别的信息特征就可以完成，甚至无法确切知道有多少个组，因此不存在用来判断模型分类效果的外部标准，聚类模型被称作不受监督学习模型。对用聚类算法构建的模型，没有对与错的答案之分，模型得出的结果由模型捕捉到的数据信息决定。

聚类方法是基于对字段记录间的距离及聚类见距离的测量，将每个记录分配给不同类别时尽量缩短属于用一个聚类记录的距离。本文便是采用聚类算法对用户的得分进行聚类分析。

4.1.1 因子分析

一、因子分析

因子分析针对多变量的情况，当变量之间有较强的相关关系的时候，用这些变量分析具体的显示问题，会使得模型因为存在多重共线性的问题而有误差。因子分析便是能有效的用新指标代替原来的多变量的方法，它们能有效的提取信息，众多的变量进行降维。

因子分析的数学模型：

$$F_1=a_{11}X_{11}+a_{21}X_{21}+\ldots+a_{p1}X_p+\varepsilon_1$$

$$F_2=a_{12}X_{12}+a_{22}X_{22}+\ldots+a_{p2}X_p+\varepsilon_2$$

…

$$F_p=a_{1m}X_{1m}+a_{2m}X_{2m}+\ldots+a_{pm}X_p+\varepsilon_p$$

$X_1,X_2,\ldots X_n$是原始变量经过数据标准化的处理，用来消除纲标影响。

系数aij是第i个变量与第j因子之间相关程度，是反映变量和因子之间相关程度，因子是出现在最初变量和因子的线性组合中，被称作是公因子，ε是特殊因子表示公因子以外的影响因素。

因子分析可以使得因子能得到更好的解释，在解释现实研究对象时，因子分析占优势，因子分析是对原始变量的信息进行了重新组合，从中找出影响变量的公共因子，进行数据的化简。

因子分析是研究最初始变量内部关系出发，把关系复杂的属性变量表示为少数的几个公因子，即要从数据中提取最具有解释性的少数公因子，因子分析和主成分分析相比较来说，因子分析倾向于描述最初变量的相关关系。

因子分析可以分为两类：

因子分析有两种形式：探索性因子分析和验证性因子分析，探索性的因子分析是

为了找出事物内部的本质，验证性因子分析是用来检验已知的特定结构是否按照预期方式进行。这两种分析都是为了找出变量间的相关关系和方差协方差间关系，相关度高的变量很可能是受同样因子的影响，相关度不高的因子可能是受不同因子的影响，因子必须要尽可能多的解释变量，每个变量在因子上都有一个因子载荷，因子的解释性由因子载荷大的变量来决定。

探索性因子分析主要是为了找出影响变量的因子个数，及变量和因子之间的相关程度，因子分析之前，不必知道存在着几个因子，它们之间关系如何。

验证性因子分析是利用先验信息，和已知因子的情况下来判断是否会按照预想的方式进行，它的主要目的是决定事前定义因子的模型的拟合实际情况的能力。

本文中用到的是探索性因子分析。

一、主成分分析

主成分分析主要是进行数据的压缩和数据的解释。主成分分析是考察多变量之间关系的一种方法，是通过从最初始的几个主要变量便能概括原来所有变量的大部分信息，并且相互之间不相关，通过来寻找众多变量的综合指标，从指标综合反映事物的某方面的特征。从综合指标中总结出恰当的解释。把原数据浓缩来解决更深一步的问题。

主成分分析的数学模型

通常用F1、F2等来标记筛选出来综合指标，综合指标被希望涵盖原变量尽可能多的信息，一般是通过方差来反映F1等综合指标的信息量，F1方差越大则代表F1包含的信息越多，在所有的综合指标中，第一个综合指标涵盖的信息应该是最大的，然后再有第二个、第三个综合指标，这些综合指标之间是互不相关，并且var（F2）<var（F1），方差递减，通常选择前面最大的几个主成分，从原始变量能提取大部分的信息。

最终得出主成分模型如下：

$F_1=a_{11}X_{11}+a_{21}X_{21}+\ldots+a_{p1}X_p$

$F_2=a_{12}X_{12}+a_{22}X_{22}+\ldots+a_{p2}X_p$

…

$Fp=a_{1m}X_{1m}+a_{2m}X2_m+\ldots+a_{pm}X_p$

其中a11,a21,…api(i=1,…,m)为X的协差阵Σ的特征值所对应的特征向量，

通过计算可以得出特征根与特征向量，有几个特征根大于1决定了选取主成分的个数，特征根小于1代表主成分的解释性还不够，特征根大于1是主成分提取的原则。通过特征根大于1的方差累计贡献来确定这几个主成分对原来变量的解释情况。

通过主成分的因子载荷得到主成分的表达式。

特征根即为每个主成分的方差，特征根越大意味着对应的主成分能够解释的原来

的信息量就越大（方差贡献率表示），累计贡献率大于85%，被用来作为提取特征值的个数的原则。

因子载荷矩阵中，每个载荷量代表主成分与对应变量的相关系数。

主成分特征向量代表主成分与相应的原先变量的相关关系，是通过载荷与特征值的平方根之比得到，绝对值越大代表变量代表性也越大。

三、因子分析与主成分分析的比较

主成分分析是通过几个主成分来解释多变量的方差-协方差的方法，主成分分析是通过求出几个主要成分，使他们保留原始变量尽可能多的信息，是把一组相关的变量转化成不相关的变量，同时在转化过程中保持变量总方差不了，具有最大方差的被称为第一主成分，具有第二大方差的，称为第二主成分。

因子分析是研究怎样用最少的信息丢失，将众多原始变量浓缩成少数几个因子变量，并且使得因子变量有较强的可解释性

分析因子分析与主成分分析的实际特点，根据实际情况的需要，本文采用因子分析来进行数据维度的降低。

4.1.2 聚类算法

聚类是将样本或者指标进行分类，是数据挖掘的一个重要算法。通过将个体或者对象分类，使得同一类中的对象之间的相似度与不同类的对象的相似度相比更强，使同类之间差异最小，异类对象之间的差异最大。

在对样本或者变量的研究中，研究对象之间可能存在着不同程度的相似性，根据样本的多个角度的观测指标，能得到一些能够度量样本或者变量相似度的统计量，把这些统计量来作为分类的依据，把相似性较大的样本或者变量划分为一类，把联系密切的样本聚合为一类，把另外一些联系稀疏的聚合到大的分类，一直到把所有的样本或者变量聚合完毕。

一、K 均值聚类算法

K 均值聚类法是一种非谱系聚类算法，非谱系聚类法是把样本聚集成 k 个类，类的个数可以提前设定或者在聚类过程中确定，它可用于比系统聚类法大得多的数据组。它是在最开始就对样本分组，从构成每个类核心的“种子”开始，可以随机的从样本中选取“种子”，随机的把样本分成若干类。

K 均值算法是基于目标函数聚类方法的一种，它的通过优化目标函数（样本点到某个点的某种距离），通过采用函数求极值方法来迭代计算，k 均值是计算对于某一个初始聚类来说的最优分类。

聚类终止的条件有两个：1）迭代次数，当建模运行的迭代次数等于制定迭代次数时终止聚类。2）类中心点的偏移程度，当偏移程度增大时停止。

样本最终的聚类较大称得上依赖初始划分或“种子”点的选择。可以通过一个新的初始分类检验聚类的效果，如果结果与原来一致，则不用再另行计算。

K 均值聚类算法缺点

（1） 当有多个“种子”跑到一个类中时，聚类的结果将比较难区分。

（2） K 均值算法中的 k 是提前就设定好的，这个 k 值得选取是不易估计的，之前并不知道所到底给定的样本量应该被分成多少个类别才比较合适，这是 k 均值算法的缺点，k 均值算法中个数的多少 k 是以方差分析理论和 F 统计量为基础来确定最佳分类数。

（3） 即使知道总体由 k 个类组成，抽样可能造成样本中不出现稀疏类的样本，强行把这些数据分成 k 类会导致无意义聚类[31-32]。

二、两步聚类

两步聚类是（两阶聚类算法）的缩写，是分层聚类算法的一种，两步聚类的特点很鲜明，进行两步聚类的变量可以是连续变量也可以是离散变量并不像 K 均值算法，在进行聚类前需要对离散变量转换为数值变量，并且运行速度较快。

在两步聚类中假设所有变量是相互独立的，两步聚类是利用似然距离来对连续变量和分类变量做处理，并且假设分类变量是多项分布的，连续变量都符合正态分布。

两步聚类分两步骤完成：

（1） 预聚类：是对样本初步归类，最大类别数可以自行指定。

第一步中是采用 BRICH 算法进行预聚类。

BIRCH 聚类算法是针对大数据集，是通过将数据以压缩的格式存放，BIRCH 是在压缩的数据集上而不是在原始数据集上聚类。BIRCH 中的两个重要概念：聚类特征（CF）和聚类特征树（CF_tree）,把簇的特征描述为这两个概念，利用层次方法对数据集根据簇之间的距离进行聚类，该方法对大规模数据处理比较鲁棒。

聚类特征

BIRCH 是通过聚类特征 CF 对每个簇所蕴含的信息进行汇总说明，再对簇进行聚类。它是通过假定一个簇中有 P 个 d 维的数据点，{Xi}，其中 i=1，2,…，N，那么聚类特征定义成一个三元组：CF=（N，LS，SS），在三元组中 N 是族中数据点数量，LS 是 N 个数据点的线性和，SS 是 N 个数据点的平方和，SS 包含了聚类和储存性能的关键信息。通常用欧几里得距离和曼哈顿距离来计算聚类特征来求出簇之间距离。

聚类特征树

BIRCH 的 CF_tree 是含有两个参数的平衡树，他存有层次聚类的簇的主要特征，CF 树的结构，根据定义，树中如果不是叶子节点，那么它将包含子节点的所有聚类

信息，它存有子节点的 CF 值总和，子节点的聚类特征和连接它的指针构成了条目，CF_tree 中包含了两个参数：分支因子 H（非叶节点包含有孩子的最大个数）和阈值 T（限制类叶节点簇的最大半径），由这两个共同决定最终数的大小，

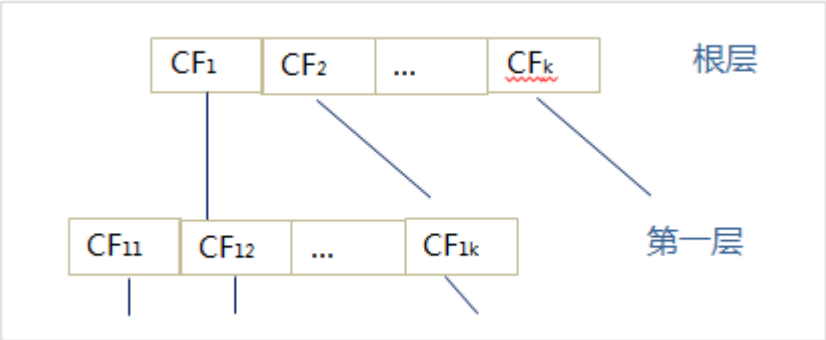

图 4.2　CF 数结构

CF 树是根据第一个样本建立起根节点好队形的条目，随后通过距离最小原则将后续样本分配到 CF 数中，经过扫面数据集不断添加更新 CF 条目和分裂点来形成 CF 树，CF 中所有节点的所有条目都代表一个子聚类。

（2） 正式聚类：对完成初步聚类的样本再聚类。

预聚类：在得到 CF 树后，所有条目都代表 1 个子聚类，然后采用凝聚的层次聚类进行再聚类。

凝聚的层次聚类是给定要聚类的 n 个子聚类，分别计算它们的 n*n 矩阵，找到比较接近的两个类合并为一个，那么总的聚类将减少一个，再重新计算新形成的聚类与所有子聚类的距离，选择较为接近的两个类进行合并，直到最终停止。

BIC（Bayesian information criterion）贝叶斯信息准则，也被称作 BIC 评分，是一种评分函数，近似与大样本数据中对边缘函数的估计，通过 BIC 准则可以自动确定最优类数。通常 BIC 是随着聚类个数增加而先减少之后增加，当 BIC 是最小的时候，被认为是最优聚类。

三、聚类算法的比较

两步聚类它的适合于处理海量数据，运算速度较快，两步聚类的一个优点在于它既能够处理同时包含离散数据和连续数据的数据集，两步聚类是通过 BIC 准则自动确定聚类数。

K 均值聚类是划分算法的一种，是用各类的所有数据的平均值所示，是较经典的算法之一，k 均值聚类算法的特点是它的效率高，k 均值算法的缺陷在于，结果的好坏很大程度上依赖于聚类中心的选择，很可能会形成局部最优，K 均值算法对异常数据较敏感，只能处理数值型的变量，聚类的结构可能不平衡。

由于 k 均值只能处理连续性变量，因此对本论文来说，两步聚类方法更适合。

4.2 使用 SPSS 及 SPSS Modeler 软件进行分析

4.2.1 数据的选取

把纳入离网倾向预测评估中的变量进行分析，从中把分类变量进行剥离，把诸如用户的城市所在，是否集团用户及七大渠道等已经固定的属性变量进行删除，在剥离固定属性的变量后得到关于用户的消费属性变量如表 4.1。

表 4.1 因子分析变量表

用户属性变量	说明
在网时长：	单位：月
前六个月平均累计欠费金额：	单位为元
前六个月平均预存款总额：	单位为元
前六个月平均出账金额：	单位为元
前六个月平均同网通话次数：	网内通话次数
前六个月平均本地主叫计费时长：	单位为分钟.
前六个月平均本地被叫计费时长：	单位为分钟.
前六个月平均上网流量：	单位：M
前三个月平均半停次数：	单位为次
前三个月平均停机次数：	单位为次
半年平均异网通话次数：	单位为次
半年平均通话次数：	单位为次

4.2.2 验证变量是否符合做因子分析

对提取的出的连续变量采用因子分析的算法来对用户的消费数据变量提取公共因子，因为这些变量中存在相关性较强的变量如图 4.3，平均异网通话次数与平均的出账金额存在较大的相关性。

		前六个月平均累计欠费金额	前六个月平均预存款总额	前六个月平均出账金额	前六个月平均同网通话次数	前六个月平均上网流量	前三个月平均半停次数	前三个月平均停机次数	前六个月平均异网通话次数	在网时长
相关	前六个月平均累计欠费金额	1.000	.045	-.074	-.051	.131	-.011	-.014	-.071	-.072
	前六个月平均预存款总额	.045	1.000	.345	.331	.323	.190	-.029	.413	-.108
	前六个月平均出账金额	-.074	.345	1.000	.646	.175	.138	-.004	.749	.094
	前六个月平均同网通话次数	-.051	.331	.646	1.000	.224	.121	-.028	.530	.000
	前六个月平均上网流量	.131	.323	.175	.224	1.000	.106	-.058	.192	-.168
	前三个月平均半停次数	-.011	.190	.138	.121	.106	1.000	.093	.191	-.123
	前三个月平均停机次数	-.014	-.029	-.004	-.028	-.058	.093	1.000	-.014	.071
	前六个月平均异网通话次数	-.071	.413	.749	.530	.192	.191	-.014	1.000	.053
	在网时长	-.072	-.108	.094	.000	-.168	-.123	.071	.053	1.000

图 4.3　变量间相关性图

通过相关性分析，得出变量间存在着很大的相关性，如图 4.3 前六个月的平均异网通话次数与前六个月的出账金额存在较大的相关性，相关系数达到了 0.749，平均同网通话次数与平均出账金额的相关性达到了 0.646 因此可以对变量采取因子分析来提取出其中的的公共因子，以得出对原来的变量比较有说明性的综合成分。

KMO 和 Bartlett 的检验

取样足够度的 Kaiser-Meyer-Olkin 度量。		.818
Bartlett 的球形度检验	近似卡方	13885399.58
	df	55
	Sig.	.000

图 4.4　KMO 和 Bartlett 的检验图

KMO 检验统计量是用来比较变量间的相关系数和偏相关系数的指标，KMO 统计量的是在 0 到 1 之间取值，如果 KMO 值越接近于 1，就意味着变量间的相关性越强，那么用原始变量做因子分析是合适的，当 KMO 统计量接近于 0 时，那么意味着变量间的相关性弱，是不适合做因子分析的。图 4.4 通过 KMO 统计量的检验得值为 0.818，较接近于 1，那么代表用因子分析来提取公因子的方法是可行的。

4.2.3 因子分析

1、数据的标化处理

因子分析就是分析变量间的关系并发现关系模式，用这几个变量就能解释所有变量，通常这样的几个变量是不可观测的。

把抽取的数据变量采用 spss 模型进行因子分析，

首先将原始数据标准化；

需要如下变换：

$$xi=\frac{Xi-\bar{x}}{s}$$

其中，xi 就是标准化之后的数据，Xi 就是变量值，$\bar{X}$ 是平均值，s 是标准差。

因子的提取方法和因子载荷的求得可以通过基于主成分的主成分分析法、基于因子分析模型的主轴因子法、极大似然法、最小二乘法、a 因子提取法、映像分析法等都可以用来得出公共因子和因子载荷矩阵。主成分分析可以为因子分析提供初始解，因子分析是主成分分析的延伸。

2、公因子的提取

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.934	39.339	39.339	3.934	39.339	39.339	3.168	31.676	31.676
2	1.496	14.958	54.297	1.496	14.958	54.297	1.943	19.427	51.102
3	1.000	10.005	64.302	1.000	10.005	64.302	1.021	10.207	61.309
4	.923	9.235	73.537	.923	9.235	73.537	1.015	10.148	71.457
5	.796	7.956	81.493	.796	7.956	81.493	1.004	10.036	81.493
6	.672	6.717	88.210						
7	.470	4.702	92.911						
8	.293	2.934	95.845						
9	.225	2.255	98.100						
10	.190	1.900	100.000						

Extraction Method: Principal Component Analysis.

图 4.5　因子特征根及方差贡献率表

上图 4.5 给出了因子特征根及方差贡献率的情况，前五个因子的累计贡献率已经达到了 81.493%，公认当因子的累计贡献率达到 80%，就可以了，因此可以认为现在提取出的四个变量已经可以很好的解释初始变量的综合信息，它涵盖了初始信息的 85%左右，具有很好的代表性。如下图 4.6 即为碎石图，是将因子与特征根的关系用图表示出来，分别显示每个因子携带的平均信息量，可以看出特征值越大代表携带的信息量越多。

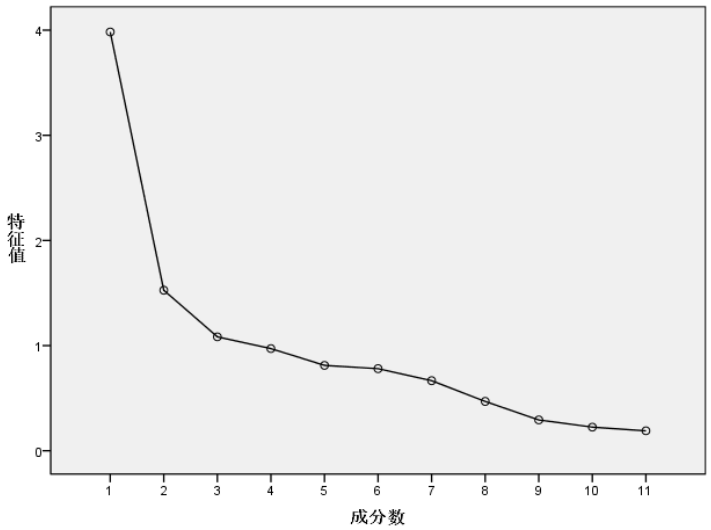

图 4.6　碎石图

3、原始变量被抽取信息量

每个变量被抽取的信息量如下图 4.7 所示，前六个月平均累计欠费金额被抽取了 99.3%，在网时长被抽取了原始变量的 98.9%，被抽取最多的三个变量是欠费金额、在网时长、平均停机次数

	Initial	Extraction
在网时长	1.000	.989
前六个月平均累计欠费金额	1.000	.993
平均预存款总额	1.000	.679
平均出账金额	1.000	.808
平均同网通话次数	1.000	.664
平均异网通话次数	1.000	.769
平均本地主叫计费时长	1.000	.827
平均本地被叫计费时长	1.000	.812
平均上网流量	1.000	.624
平均停机次数	1.000	.985

图 4.7　初始变量被抽取的信息比例

4、旋转后的成分矩阵

Rotated Component Matrix(a)

	Component				
	1	2	3	4	5
在网时长	.045	-.133	.984	-.016	-.028
前六个月平均累计欠费金额	-.057	.086	-.027	-.004	.991
平均预存款总额	.290	.762	.010	.107	-.047
平均出账金额	.427	.768	-.163	.080	.048
平均同网通话次数	.800	.141	-.052	-.030	-.002
平均异网通话次数	.834	.243	.073	.083	-.052
平均本地主叫计费时长	.877	.233	-.041	.044	-.031
平均本地被叫计费时长	.887	.119	.096	.031	-.026
平均上网流量	.045	.773	-.077	-.076	.113
平均停机次数	.052	.038	-.016	.990	-.004

图 4.8　旋转成分矩阵图

进行因子分析之后，新得出的因子是原始变量的线性组合，它具有比原始变量更强的解释能力。经线性组合可表示为：

假设第一列的变量分别设为前六个月平均本地主叫计费时长为 X_1，前六个月平均本地被叫计费时长为 X_2，前六个月平均异网通话次数为 X_3，依次类推，前三个月平均半停次数为 X_{11},那么所得出的公因子 F_1、F_2、F_3 可以表示为：

$F_1=-0.045X_1-0.57X_2+0.29X_3+0.427X_4+0.80X_5+0.834X_6+0.877X_7+0.887X_8-0.045X_9+0.052X_{10}$

$F_2=-0.133X_1-0.086X_2+0.762X_3+0.768X_4+0.141X_5+0.243X_6+0.233X_7+0.119X_8-0.773X_9+0.038X_{10}$

。。。

$F_5=-0.028X_1+0.991X_2-0.47X_3-0,048X_4-0.002X_5-0.52X_6-0.031X_7-0.026X_8+0.113X_9-0.004X_{10}$

表 4.2 公共因子解释说明表

重新命名	提取的公共因子	原始变量
用户通话消费	F1	平均同网通话次数
		平均异网通话次数
		平均本地主叫计费时长
		平均本地被叫计费时长
用户消费及上网情况	F2	平均预存款总额
		平均出账金额
		平均上网流量
在网时长	F3	在网时长
停机次数	F4	平均停机次数
欠费情况	F5	前六个月平均累计欠费金额

把因子分析得出的公共因子可以概括为解释性较强的变量，如表 4.2 中 F1 的公共因子可以归纳为用户的通话消费情况，F2 的公共因子可以被归纳为用户总消费情况，而剩变量的原因是因为因子分析的原则是要抽取出的变量要包含所有原始信息量的 80%。

当因子分析结束后，可以得出每个因子对总因子的相对重要性用方差贡献率来衡量，把方差贡献率的多少来作为权重，计算公共因子中各个因子的相对重要性，可以作为得分来反映一个用户综合的消费情况。

权重表如下：

表 4.3 因子得分表

	F1	F2	F3	F4	F5	Σ
贡献率	39.339	14.958	10.005	9.25	7.956	81.493
权数	0.482	0.183	0.122	0.11	0.103	1

通过以上表 4.3 因子得分表可以得出公共因子之间的相对重要性。

通过以上分析可以得出因子分析的结果，并保存公共因子为新变量，4.2.3 是把用户的消费行为变量进行聚类，得出主要的可解释较强的五种因子，如用户通话消费、用户消费及上网情况、用户的在网时长、停机次数、欠费情况五种可解释新变量。为下一步的聚类分析做准备。

4.2.4 两步聚类分析

本小节是把经过上一小节提取的公共因子作为用户的消费指标，来和用户的其他固定属性变量综合起来进行聚类分析。

其中用户的固定属性变量包括（如表 4.4）：

表 4.4　用户固定属性变量表字段

用户属性变量	说明
USER_NO	用户的唯一识别
是否集团	1 是、2 否
合约类型：	存费送机.

把因子分析得出的五个公共因子与用户的两个个属性变量进行两步聚类。UER_ID 作为用户的唯一识别标示，但是不参与聚类。

其中是否集团为名义变量，定义它的取值为 0 或 1，合约类型是包括普通单卡、存费送机、存费送业务、购机入网和其他。把这些字段定义为有序变量 1、2、3、4. 在因子分析数据流的基础上继续进行聚类分析：

1、采用 spss modeler 进行两步聚类的数据流

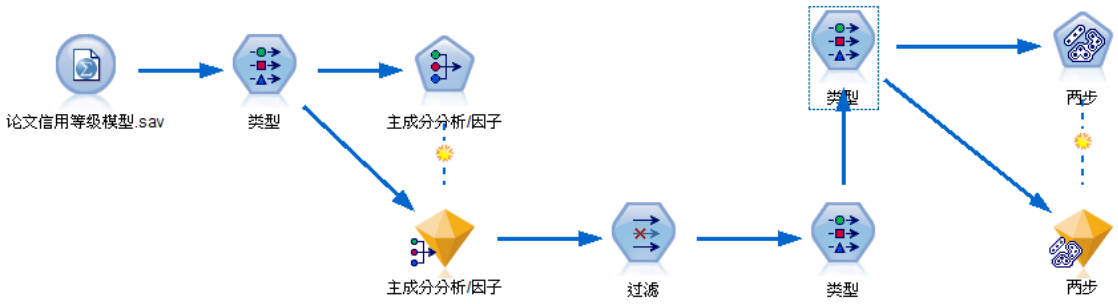

图 4.9　聚类分析数据流图

因子分析得出的新变量与表用户固定属性变量表字段公共作为聚类的变量。

2、聚类结果展示

表 4.5 聚类结果表

聚类-1	聚类-4	聚类-3	聚类-2
因子-1 -0.18	因子-1 -0.04	因子-1 0.11	因子-1 0.65
因子-2 -0.19	因子-2 -0.25	因子-2 -0.23	因子-2 1.01
因子-3 0.16	因子-3 0.37	因子-3 -0.69	因子-3 -0.43
因子-4 0.03	因子-4 0.02	因子-4 -0.13	因子-4 0.06
因子-5 -0.04	因子-5 -0.03	因子-5 -0.03	因子-5 0.28
合约类型 普通单卡（100%）	合约类型 普通单卡（100%）	合约类型 存费送机（86.8%）	合约类型 存费送机（85.6%）
是否集团 0（100%）	是否集团 1（100%）	是否集团 0（100%）	是否集团 1（91.5%）

表 4.5 中显示聚类结果中每个因子的分布状况。

通过聚出的 4 类，进行结果分析。并且结合因子 1（社会网络关系因子）、因子 2（消费娱乐因子）、因子 3（在网时长因子）、因子 4（停机次数因子）、因子 5（欠费因子）来进行综合分析。聚类结果应该与现实情况相结合，判断出每一类用户的行为特征，进而再对每一类用户授信，通过聚类结果得出的分类进行授信具有一定的主观性，但仍不失为一个为用户授信的良好办法。

4.3 消费异动价值评估

消费异动价值评估是动态信用评估的第三部分。

消费异动价值评估主要是分析通话时长（主叫计费时长）和用户办理新的业务（包括产品订购信息、短信详单、通话详单、流量信息这些变量综合起来考虑）等异常消费变动对停机造成的影响。

定义：风险因子（新办业务）=消费异动类型造成停机的数量/样本中总的消费异动类型的数量。

举例说明：如果办理某项套餐的用户为 100 万人，办理这项套餐的用户中有 40

万停机，那么该项套餐业务的风险因子即为 40/100=0.4.通过这样的方式把每项套餐的特殊情况考虑在内，当办理某项套餐而停机的用户数量较少时，那么该项套餐业务的风险因子就较小，当办理某项套餐而停机的用户数量较多时，那么该项套餐业务的风险因子就会较大。

发生主叫通话时长消费异动的判断方式为：（主叫计费时长>=平均值+2*标准差），当月订购的增值业务产品也判断为本月的消费异动。

消费异动价值（异常主叫）：W0=0.35*（主叫计费时长-平均值-2*标准差）；

消费异动价值（新办业务）：$W1=k_1*h_1*X_1+k_2*h_2*X_2+ \cdots\cdots +k_n*h_n*X_n$；

X_i 为消费异动类型，k_i 判断 X_i 是否发生，如果 X_i 发生，k_i 取值为 1，如果 X_i 不发生，k_i 取值为 0，h_i 为消费异动类型 X_i 的风险因子；

总的消费异动价值=消费异动价值（异常主叫）+消费异动价值（新办业务）

按照当前的增值业务产品计算的各增值业务产品的风险因子如下：

表 4.6　增值业务风险因子表

增值业务产品 ID	增值业务产品名称	产品订购数量	用户停机数量	停机比率（**风险因子**）
		14453906	3522424	0.24
47558	手机邮箱 3G 版	4574797	961718	0.21
75043	国内流量充值 30 元包 400M(当月有效)	1802555	380606	0.21
70418	3G 国内流量包-10 元套餐	711057	174324	0.25
70414	3G 国内流量包-30 元套餐	426841	118399	0.28
75044	国内流量充值 50 元包 800M(当月有效)	410887	85690	0.21
68571	本地炫铃功能_5 元/月	408535	111466	0.27
70413	3G 国内流量包-20 元套餐	365259	89781	0.25
69993	省内流量 10 元包 100M(本省)	336175	94076	0.28
75042	国内流量充值 20 元包 200M(当月有效)	309217	77423	0.25
64268	GPRS5 元包 30M(500 元、6G 双封顶)	182086	49406	0.27
69994	省内流量 20 元包 300M(本省)	136558	41416	0.30
74110	本省流量 5 元包 500M（夜间 23 点-8 点优惠包）	126628	34862	0.28

77542	省内流量 5 元包 60M(本省)	107063	25566	0.24
74113	本省流量 10 元包 1G（夜间 23 点-8 点优惠包）	90188	25657	0.28
64750	GPRS2 元包 6M(500 元、6G 双封顶)	63913	16387	0.26
64753	GPRS10 元包 70M(500 元、6G 双封顶)	54880	15542	0.28
70813	省内流量 30 元包 500M(本省)	43484	14808	0.34
74114	本省流量 20 元包 3G（夜间 23 点-8 点优惠包）	20117	5356	0.27

表 4.5 为部分业务套餐风险因子的示例。

通过以上表格得到了每种增值业务产品所对应的风险因子，这样便可计算出每个用户的总消费异动价值为异常主叫消费异动价值与新办业务消费异动价值之和共同构成。

消费异动价值是通过衡量用户的实时行为动态来设定实时的用户信用额度，这样会把根据每个用户的特殊情况而赋予用户相匹配的信用额度，在现实应用中，更加合理，用户体验会更好。

4.4 本章小结

本章是对用户信用等级评估与消费价值异动模型做了总结分析，用户信用等级评估是采用了 k 均值聚类算法，以 A 公司内部的属性打分的规则对用户根据属性变量计算得分，在根据每个用户得分进行分类，如果采用系统聚类法，则由于庞大的用户数据，要对 n 个样本分成 n 类，进行繁杂的计算，然而针对 k 均值聚类通过提前设定好分类数，本文中是提前设定好分类数为 5，再对现有用户数据集进行分类，使得操作简洁，得到有效的用户分类，进而赋予每个类别一定的信用额度。

消费异动价值评估是为了把用户的实时行为的变化考虑进去，消费异动价值是通过两个部分来构成的，一部分是异常主叫影响，另一部分是新业务办理的异动价值。当用户为正常消费用户，某个月主叫时长突然增加，会对这类客户赋予更多一些的信用额度，以保证他们不停机。当正常用户办理新业务时，会对用户赋予新办业务同等价值的信用额度，以保证多消费会享有更高的信用额度。通过这两种消费异动价值变化来赋予用户不用的信用额度，保证信用额度的差异化来对用户提供较好的服务。

第五章 离网预测与信用评价结果评价

第三章的离网倾向预测与第四章的用户信用等级和消费异动价值共同构成了本文针对 A 公司的用户的动态信用评级。

离网倾向预测与信用评价结果主要从以下几方面展开：离网倾向预测模型的结果及评估、用户信用等级模型的结果及评估、用户信用度的构成及和 A 公司现有系统的信用度模型的比较分析。

1.离网倾向预测的结果

离网倾向预测的作用是形成对高危离网用户的判断准则，通过该判断准则我们可以通过将未知用户的属性数据纳入判断准则，从而判断用户是正常用户还是高危用户。

离网倾向预测的规则集如下（判断用户为高危离网用户的规则，不满足以下规则的任一条则判断该用户正常用户，由于篇幅的限制，只列举出前 5 个判断规则）：

规则 1

如果在网时长> 1

和合约类型 = 购机入网

和半年平均累计欠费金额<= 113.062

和半年平均出账金额<= 42.835

和 06 月出账金额离均差> 12.187

和半年平均异网通话次数<= 6.667

规则 2

如果在网时长> 16

和在网时长<= 18

和合约类型 = 存费送机

和半年平均累计欠费金额> 47.917

和 06 月累计欠费金额离均差> 0.005

和 06 月累计欠费金额离均差<= 11.880

和 06 月出账金额离均差> -8.175

规则 3

如果在网时长> 1

和合约类型 = 购机入网

和半年平均累计欠费金额<= 113.062

和半年平均出账金额<= 42.835

和 06 月累计欠费金额离均差> -17.655
和 06 月出账金额离均差> 18.383
和 06 月本地被叫计费时长离均差<= -1
和 06 月半停次数离均差<= 0.333
和半年平均异网通话次数<= 70
规则 4
如果在网时长> 1
和半年平均累计欠费金额> 15.470
和 06 月累计欠费金额离均差> 11.880
和 06 月出账金额离均差> 0.733
和 06 月上网流量离均差<= 0.000
和 4 到 6 月平均停机次数<= 0
和半年平均通话次数> 0
规则 5
如果在网时长> 1
和合约类型 = 购机入网
和半年平均累计欠费金额<= 113.062
和半年平均出账金额> 42.835
和半年平均出账金额<= 48.315
和 06 月累计欠费金额离均差> -42.418
和 06 月出账金额离均差> 17.050
和 06 月停机次数离均差<= 0
和半年平均通话次数<= 45.667
规则 6 用于 1 (219; 0.814)
如果 在网时长 > 1
和 合约类型 = 购机入网
和 半年平均累计欠费金额 > 42.593
和 半年平均出账金额 > 42.835
和 06 月累计欠费金额离均差 > -42.418
和 06 月累计欠费金额离均差 <= 11.880
规则 7 用于 1 (830; 0.799)
如果 在网时长 > 1
和 合约类型 = 购机入网
和 半年平均累计欠费金额 > 113.062

和 06 月出账金额离均差 > -2.100

规则 8 用于 1 (2,318; 0.799)

如果 合约类型 = 购机入网

和 半年平均出账金额 > 45.107

和 06 月累计欠费金额离均差 > 11.880

和 06 月出账金额离均差 > -44.785

和 4 到 6 月平均停机次数 <= 0

规则 9 用于 1 (377; 0.728)

如果 地市 = 宝鸡

和 在网时长 > 19

和 在网时长 <= 21

和 合约类型 = 存费送机

和 半年平均累计欠费金额 <= 12.648

和 06 月停机次数离均差 > 0.333

和 6 月异网通话次数离均差 > -169.833

规则 10 用于 1 (4,833; 0.668)

如果 在网时长 > 1

和 合约类型 = 购机入网

和 06 月累计欠费金额离均差 > 11.880

和 06 月出账金额离均差 > 0.733

以上为列出的规则集的示例，判断为用户为高危离网用户的规则总共有 65 条。每个规则集均包含对多种变量的判断条件。

不同的变量对判断用户是否为高危离网用户的重要性是不一样的，下图列出了在预测用户是否为高危用户的变量的重要性的排序结果。变量重要性的确定方法为：(1-Pi)/sum(1-Pi)，其中 pi 是该变量与目标变量之间的显著性检验结果值，i 为变量的个数；从该图中我们可以发现上月本地被叫计费时长离均差、七大渠道、上月上网离均差、是否沃家庭和前六个月平均预存款总额这五个变量是对于模型的构建相对重要的变量（如图 5.1）。

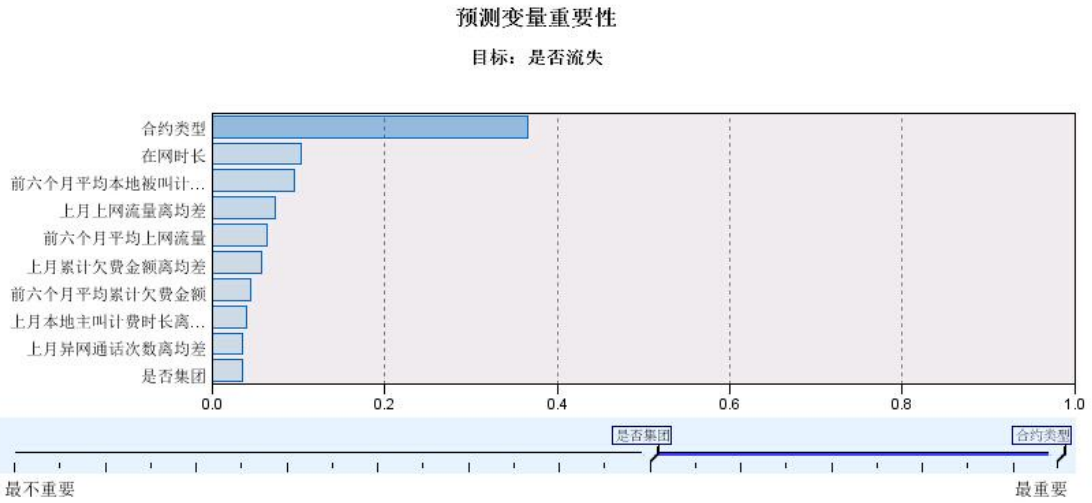

图 5.1　预测变量重要性图

2.离网倾向预测的评估

离网倾向预测的结果主要从重合矩阵、增益图和提升图来进行评价。

重合矩阵又被称为混淆矩阵，适用于分类型输出变量，其中行数据是实际值，而列数据表示预测值，此矩阵能够告诉我们模型分别对正常用户和高危用户的的预测效果。

根据预测值及预测的置信度排序记录、将记录分割为大小相等的组（分位数）并按由高到低顺序为每个分位数绘制业务标准值，增益的定义是相对于全部匹配，发生于每个分位数中的匹配的百分比。其计算方法为（分位数中的匹配数量/全部匹配数量）×100%；累积增益图的线从左至右的走势通常是从 0% 到 100%，优秀模型的收益图将陡升至 100%，然后保持平直；无法提供有用信息的模型将呈对角线状，即从左下角到右上角（选择了包含基线后将显示类似图表）。

提升是将每个分位数中匹配记录的百分比与在全部训练数据中匹配的百分比进行比较，其计算方式为（在分位数中的匹配/在分位数中的记录）/（全部匹配/全部记录）；累积提升图的线从左至右的走势通常为：起始于大于 1.0 的值，并渐渐下降，直到接近 1.0。图表的右侧边缘表示整个数据集，因此累积分位数的匹配与数据中的匹配的比例为 1.0。对于优秀模型的提升图，其线开始于图表左侧大于 1.0 的值，且在向右移动的过程中，始终保持在较高的水平；然后，在图表右侧，向 1.0 的方向迅速下降。如果模型不能提供任何信息，则其线在整个图形中将始终围绕在 1.0 左右。如果选择了包含基线，一条值为 1.0 的水平线将显示在图表中供您参考。

离网倾向预测的混淆矩阵、增益图和提升图分别如下所示（由于建模过程中采用了抽样技术改变了高危离网用户的占比，对训练集的预测结果没有意义，因此我们对模型的评估只针对测试集进行）：

表 5.1 混淆矩阵表

测试集	预测高危用户	预测正常用户
实际高危用户	9079	1650
实际正常用户	70479	1594066

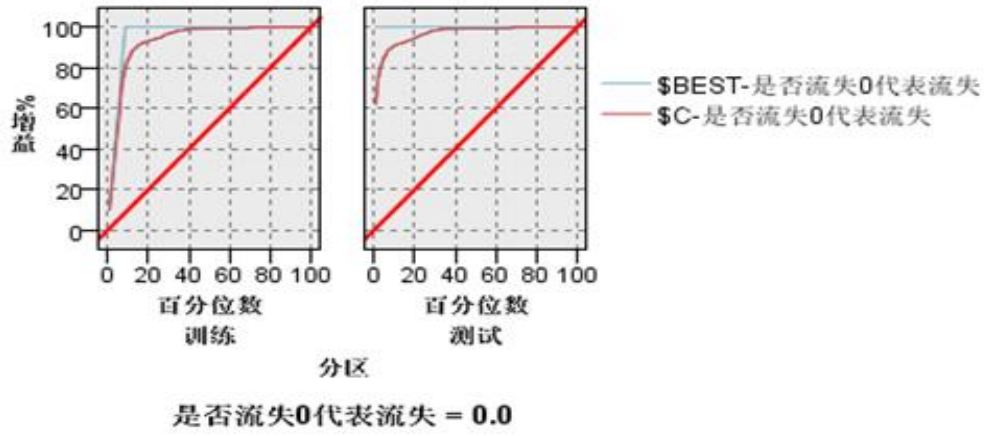

图 5.2 增益图

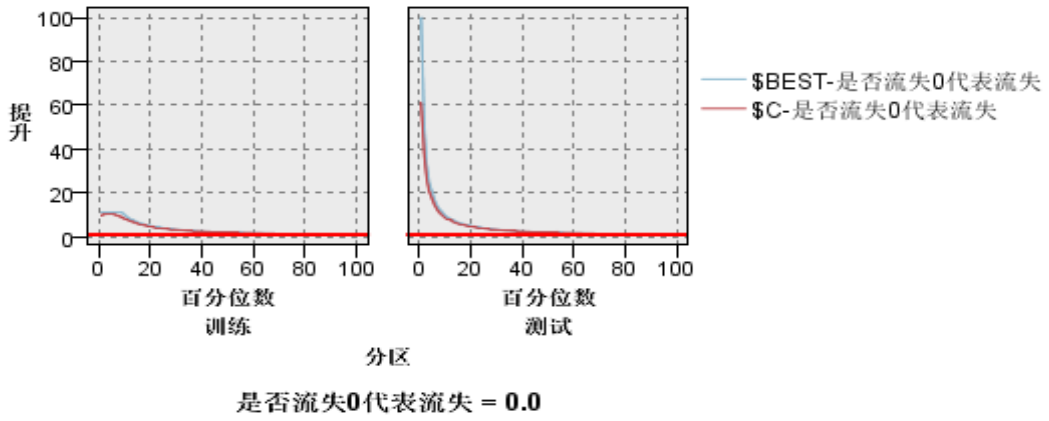

图 5.3 提升图

混淆矩阵（表 5.1）、增益图（图 5.2）和提升图（图 5.3）表明，采用 C5.0 算法对高危用户起到了比较好的分类效果；下面我们从离网用户的命中率和提升度这两个方面来比较采用 C5.0 算法和 A 公司现有离网预测模型的效果。

提升度衡量的是与不同模型相比，模型的预测能力变好了多少，对提升度结果来说，提升度指数越高，代表模型的运行效果越好，在此文中，提升度的计算公式：

Lift=预测为离网用户的人数中实际离网用户所占的比率/所用用户中离网用户所占的比率；

根据公式计算得出的模型的对高危离网用户的提升度为：

预测为离网用户的人数中实际离网用户所占的比率=9079/(9079+70479)=00.11411；

所有用户中离网用户所占的比率=(9079+1650)/(9079+1650+70479+1594066)=0.0064；

因此得出：提升度=0.11411/0.0064=17.829；

命中率是指判断为高危离网用户占实际高危离网用户中的比例，采用 C5.0 算法得出的对高危离网用户的命中率为：

命中率=9079/(9079+1650)=84.6%

A 公司现有模型的提升度和命中率分别为：8.0 和 10%，具体对比结果如下表：

表 5.2　效果评估表

	提升度	命中率
现有模型	8	10.2%
C5.0 算法建立的模型	17.8	84.6%

从表中我们可以发现，采用 C5.0 算法建立的高危离网模型要好于 A 公司现有的模型。

2.用户信用等级模型的结果

用户信用等级模型是选择与形成用户信用等级有关的变量采用聚类算法将相似的用户聚成一类，然后比较各类别属性和消费行为上差异，通过专业的知识对聚类所形成的不同的类别赋予相应的信用等级。

下图为在聚类过程中采用主成分分析方法提取的公共因子，从图中我们可以很清楚的发现 5 个因子可以得到很好的解释。

因子 1 中平均同网通话次数、平均异网通话次数、平均本地主叫计费时长、平均本地被叫计费时长这几个所占比很大，该因子可以描述为社会网络关系因子。

因子 2 中预存款总额、平均出账金额和平均上网流量这几个因子所占比很大，因此，该因子可以描述为消费娱乐因子。

因子 3 主要就是由在网时长构成，可以直接看成是在网时长因子。

因子 4 主要由平均停机次数构成，可以直接看成是停机次数因子。

因子 5 主要是由平均累计欠费金额构成，可以直接看成是欠费因子。

	Component				
	1	2	3	4	5
在网时长	.045	-.133	.984	-.016	-.028
前六个月平均累计欠费金额	-.057	.086	-.027	-.004	.991
平均预存款总额	.290	.762	.010	.107	-.047
平均出账金额	.427	.768	-.163	.080	.048
平均同网通话次数	.800	.141	-.052	-.030	-.002
平均异网通话次数	.834	.243	.073	.083	-.052
平均本地主叫计费时长	.877	.233	-.041	.044	-.031
平均本地被叫计费时长	.887	.119	.096	.031	-.026
平均上网流量	.045	.773	-.077	-.076	.113
平均停机次数	.052	.038	-.016	.990	-.004

图 5.4　成分得分系数矩阵

采用两步聚类所形成的聚类结果如下图所示，从聚类中我们可以发现总共聚成了 4 类，每类占比分别为：

表 5.3　类别占比表

类别 1	类别 2	类别 3	类别 4
43.2%	27.5%	18.3%	11.0%

类别 1 中各变量的数值中我们可以发现该类的用户在各特征变量均不明显，可以将该用户看成是普通用户。

类别 2 中因子 3 比较突出，及该类用户的在网时长比较长，并且均为集团用户，因此该类可以看成是具有一定的在网时长的普通集团客户。

类别 3 中用户的社会网络关系和消费娱乐因子均比第一类和第二类要大，但是在网时长要短，并且合约类型大部分为存费送机。

类别 4 中用户的社会网络关系因子，消费娱乐因子均很大，欠费因子也很大，并且都为集团用户，因此该类用户可以看成是有高活跃度的集团客户。

从类别 1 到类别 4 中不同类别的用户的属性及消费特征和所处行业的专业知识，可以很清楚的发现从类别 1 到类别 4，用户的价值是逐渐增大的，因此其信用等级也应该逐渐提高，如果信用等级按照从高到低分为 4 级，分别为 A、B、C、D，那么不同类别的用户对应的信用等级对应关系如下表：

表 5.4　类别评级表

类别	类别 1	类别 2	类别 3	类别 4
信用等级	D	C	B	A

3.信用等级模型结果的评估

聚类结果的评估通常采用两种方式，一种是结合专业知识检查聚类的结果是否合理，第二中就是检查轮廓系数。按照通信行业实际的情况，应该是普通用户占大多数，高端用户占比相对较小，从个类别的占比中，可以很清楚的发现采用两步聚类的结果和实际很符合。

轮廓系数（Silhouette Coefficient），是聚类效果好坏的一种评价方式。它结合内聚度和分离度两种因素。可以用来在相同原始数据的基础上用来评价不同算法、或者算法不同运行方式对聚类结果所产生的影响。

轮廓系数的计算方法如下：

1、对于第 i 个点，计算它到簇中所有点的平均距离为 ai。

2、对于第 i 个点和不包含该点的簇，计算该点到给定簇中所有点的聚类，对于所有的簇，找出最小值记作 bi。

3、第 i 个点的轮廓系数即为：Si=(bi-ai)/max(ai,bi)

Silhouette 的系数是 1 表示所有的点直接位于它的聚类中心上。值为-1 表示所有的点位于其他聚类的聚类中心上。值为 0 表示点到其自身聚类中心与到最近其它聚类中心是等距的。可以通过轮廓系数快速的检查聚类的质量如何，将所有的点的轮廓系数求平均，就是该聚类结果总的轮廓系数，本文中的轮廓系数在 0.9（如图 5.5），说明聚类的效果很好。

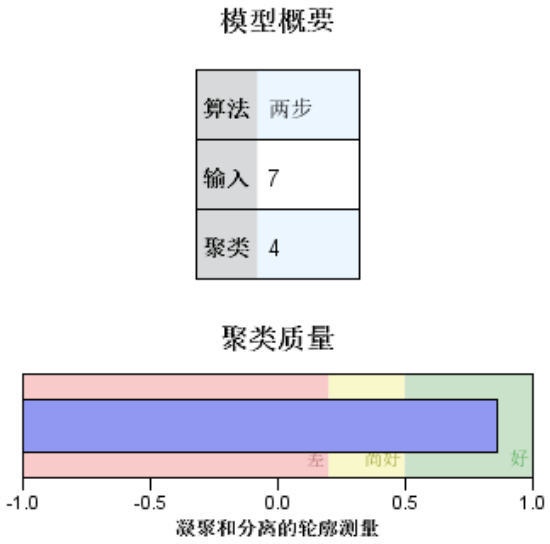

图 5.5 轮廓测量图

（1）现有的打分规则

A 公司现有的对用户分信用等级所采用的办法为将用户分成三个信用等级，信用等级的确定采用如下打分规则确定（表 5.5）：

表 5.5 信用等级打分表

是否集团	是	否					
得分	200	0					
合约类型	存费送业务	存费送机	购机入网	普通单卡			
得分	230	260	200	0			
最近 6 个月平均每月出账金额（元）	>=800	[400, 800)	[200, 400)	[100, 200)	[0, 100)	其它	
得分	200	180	160	140	60	0	
在网时长（月）	>30	[24, 30)	[18, 24)	[12, 18)	[6, 12)	<6	
得分	300	200	150	100	60	0	
最近 6 个月停机次数	0	1	2	3	4	5	>=6
得分	100	0	-100	-200	-350	-500	-700

具体的确定用户信用等级的方法是按照打分规则，计算每个用户的总得分，得分小于 200 的为等级 1，得分在 200 到 700 之间的为等级 2，大于 700 的为等级 3，为等级 1 的赋予 0 信用额度，为等级 2 的赋予 2.5 倍信用额度，为等级 3 的赋予无穷信用额度，

采用该方式所算得的各等级用户的占比如下表所示：

表 5.6　等级占比表

等级 1	等级 2	等级 3
29.5%	62.2%	8.3%

从表 5.6 中我们可以发现等级 2 的占比居然达到了 62.2%，意味着 A 公司会为大部分人赋予 2.5 倍的信用额度，这明显存在着问题（没有将用户更好的区别开），原因是该方法采用打分的办法，主观性太强，并且用户的信用等级只考虑了 5 个变量，没有综合的考虑用户其他属性变量对用户信用等级的影响，

其中的授信额度是根据用户享有的信用额度（与套餐月租费的比率）的确定是遵循这 2 条原则：1、用户的得分越高享有的信用额度（与套餐月租的比率）是越高的。2、由于本文的数据采用的是联通 3G 后付费的用户数据进行评估，3G 后付费用户的出账是采用月结的方式，因此起始信用额度（0 信用额度除外）至少要达到一个月的出账金额以上。这也是起始选择 1.5 倍原因，具体情况可以根据不同地区情况具体调整，比率的分别选择 1.5 倍，2.5 倍，3.5 倍，4.5 倍及无限就是根据 A 公司具体情况在建模过程中与 A 公司市场部、客服部讨论得出。根据业务一线的实际情况来确定信用额度的设定更有说服力。

（2）采用两步聚类算法评级的优越性

采用两步聚类算法，将用户的属性和消费行为均纳入考虑，通过用户的属性和行为划分为不同的类别，从而对不同的类别（具有不同的属性和消费行为）确定信用等级，这种方式更具有内在的合理性以及科学性。

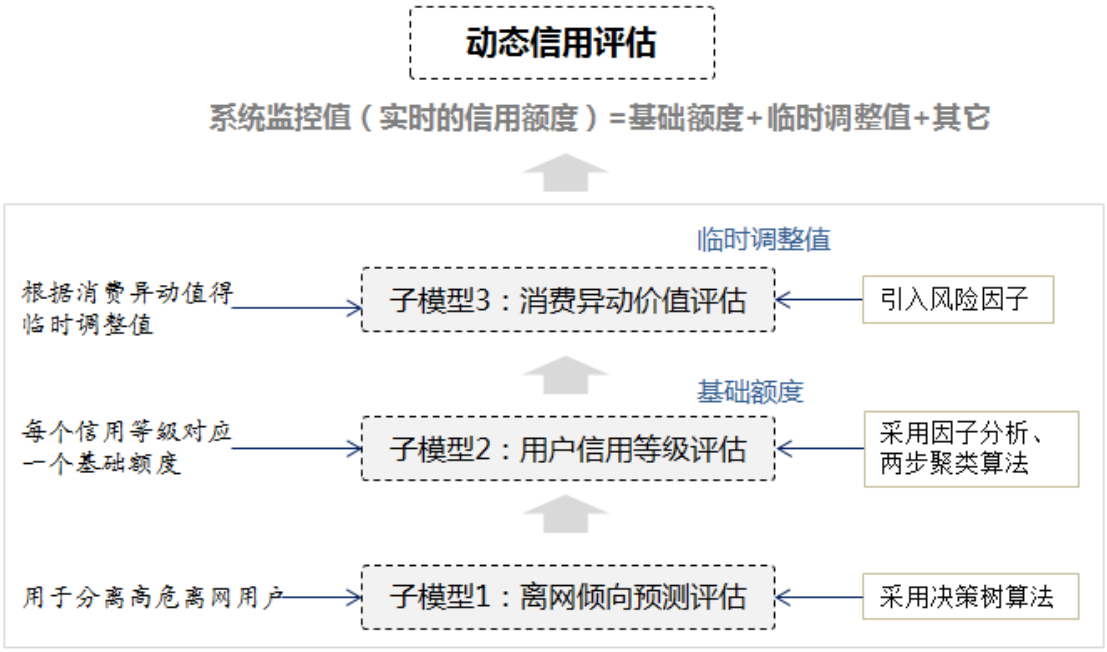

图 5.6　动态信用评估步骤图

由此得出的每个用户的信用额度综合了基础额度与临时调整值，那么实时动态信用额度由这两个部分得出的信用额度与其他服务的专属额度（如 VIP 用户享有的特殊额度等）共同得来如图 5.6。

实时信用额度=基础额度+临时调整值+其他（专属额度等）

注：专属额度指的是 VIP 用户、沃家庭用户等的专享额度，用户的信用额度是可以叠加使用。

通过用户的历史数据建立离网倾向模型可以对用户在将来是否离网做作出预测，动态信用模型的子阶段 1 即离网倾向预测评估主要是为了将用户进行分类，做到将高危离网用户进行分离。离网倾向预测评估是采用 C5.0 决策树算法，C5.0 算法较其他决策树算法 CHAID、CART 相比，算法的优势更明显，对于建模的效果更好。

用户信用等级评估是在利用现有打分的规则基础上，可以根据每个用户的属性变量（如是否为集团用户、合约类型、在网时长）各自被设定的分值等来得到每个用户的得分，得到每个用户的得分后用聚类算法进行聚类，分为 5 类，再分别对对每个级别的用户赋予一定的信用额度，信用额度的大小取决于每个用户的通过规则的得分和用户所选取得套餐额度。

消费异动价值评估是把用户的实时的通话时长和新办业务加以考虑，当用户的通话主叫超过该用户的平均值与 2 倍方差之和时，被判定为通话异常，赋予该用户异常主叫的消费异动价值，当用户办理新业务时，也会赋予用户新业务办理的消费异动价值，异

常主叫的消费异动与新业务办理的消费异动共同构成了用户的消费异动价值。

通过以上三个模型的综合，便得出了每个用户的实时动态信用评估。每个用户可以根据自己的消费历史情况与消费异动的变化而获得实时的动态信用额度。

5.1 动态信用分析的评估

5.2.1 动态信用的方法评估

对动态信用评估的效果评估也是通过对三部分的效果评估得到的。由于消费异动价值评估的效果的评估没有标准化的指标来对比，本文中主要侧重对离网倾向预测评估效果的评估和对用户信用等级效果进行评估。

1、对离网倾向预测的评估

离网倾向预测评估的建立是采用的 C5.0 决策树算法，对离网倾向预测评估的效果评估需要通过对其他的决策树算法来比较得出准确率的高低，第三章只是从定性的角度分析了 C5.0 用来建模的优势与长处，本章中用 SPSS Modeler 来对不用的决策树算法的效果做现实的印证。

本章将分别采用 CART、CHAID、QUEST 算法等与 C5.0 算法的建模结果情况做对比。

利用 CART、CHAID、QUEST 三种算法分别进行决策树建模计算，对三者建模的效果直接进行比较如下表所示：

表 5.7　四种算法建模结果准确率比较

方法	准确率
CART	93.86%
CHAID	92.3%
QUEST	94.6%
C5.0	95.11%

表 5.7 中显示采用不同决策树建模的预测准确率，均为测试集准确率，训练集被用来建模，我们更关心的是测试集的测试效果，在利用其它方法建模的过程中，采用的步骤均与采用 C5.0 建模的步骤是一致的。最终的结果通过对比显示 C5.0 的算法准确率是最好的，达到了 95.11%，其它三种算法的预测准确率也达到了不错的效果，但是相比较来说 C5.0 的算法略优。

表 5.8 四种算法建模把高危用户预测为正常用户数量表

算法	把高危预测为正常用户的数量
CART	6850
CHAID	5858
QUEST	3614
C5.0	1718

表 5.8 中为把高危用户预测为正常用户的数量表，把预测数量的大小作为衡量该模型优劣的一个标准，因为对于 A 公司实际情况来说，把高危用户预测为正常用户的数量所带来的损失是巨大的，远远大于把一个正常用户预测为高危用户带来的损失，因此宁可牺牲其他结果的准确性，也要尽量降低把高危用户预测为正常用户的数量，因此在表 5.2 中给出了各个模型用到的算法计算出的把高危用户预测为正常用户的数量做比较，并且这个量要尽可能的小，在该表中，C5.0 预测的结果 1718 为最小。因此，C5.0 算法在预测方面也是相对于其他算法更优。

综上所述，C5.0 算法是对离网倾向预测评估建模最优的决策树算法。

2、对用户信用等级方法效果评估

在因子分析中的评估主要是通过 KMO（Kaiser-Meyer-Olkin）检验模型和巴特利特球检验来对统计量来检验，在进行因子分析时，KMO 统计量为 0.818，大于 0.5，KMO 统计量的值接近于 1，巴特利特球是从变量间的相关性考虑，是服从卡方分布的，用这个度量的卡方统计量为 13885399.58，显著性概率是 0，说明变量间是相关有关联的，意味着做因子分析的是较适合的。

5.2.2 动态信用效果评估

集团内部对高危用户预测与本文离网倾向模型对高危用户预测的效果比较

集团内部现采用的是线性回归模型来对离网高危用户进行预测，现有的通过 logistic 线性回归模型对高危用户的预测效果较差，预测准确率较差，仅为 10%。

按照现有的模型规则来预测对分公司造成了很多坏账无法处理，而这部分坏账正是由这些高危离网用户造成的，因此该模型被分公司所摒弃，它远没有达到高效预测出高危用户的结果。

本文采用的决策树算法的 C5.0 算法来构造离网倾向预测评估，模型达到了很好的拟合显示数据的效果，准确率达到了 95%。并且误判损失也相对较小，把流失用户预测为正常用户的数量相对与本文建模的 340 万条数据也相对较小，本文的离网倾向预测评

估达到了不错的预测效果。

集团内部现用的信用评估规则与本文的动态信用评估比较

集团内部的 3G 后付费信用管理分册把第一版的分为未评级与 ABCD 四个级别缩减到第二版的三个级别（如图 5.7）。

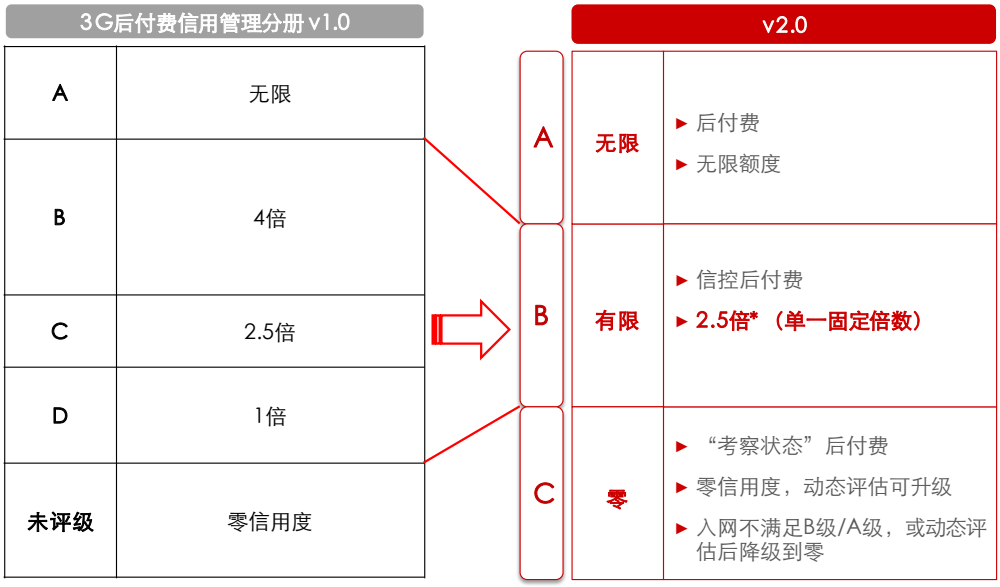

图 5.7　现有信用等级划分图

第二版的信用管理把用户等级分为 A 级、B 级、C 级三类，其中 A 类信用等级享有的信用额度是无限的，B 级信用等级享有的信用额度是有限制的，是通过公式来计算得到的，它是从第一版的五个级别变更为第二版的三个级别，第一版的 A 级别仍然为第二版的 A 级别，相应无限的信用额度，第一版的 B、C、D 级别转化为第二版的 B 级别，第一版的未评级被设定为第二版的 C 级别，信用额度授予为 0。

图 5.8 与图 5.9 为第三版信用额度的设定标准。初始入网固定信用额度与动态评估的固定信用额度分别设定为三级。均可享受 0 至无限的信用额度。在初始入网固定信用额度的设定中，被授予为 A 级的客户享有无限的信用额度，为 B 级额用户可享有 w 的信用额度：

W=套餐月费（基础套餐、叠加包等所有月租费之和）*2.5

C 级享有的信用额度为零。

在动态评估固定信用额度中，被授予为 A 级的客户享有无限的信用额度，为 B 级额用户可享有 w 的信用额度：

W=max(前 6 个月月均消费额套餐月费)*2.5

C 级享有的信用额度为零。

信用等级	初始入网固定信用额度	
	可授予额度	用户自调范围
A级	无限	0-无限
B级	W= 套餐月费（基础套餐、叠加包等所有月租费之和）*2.5 R= w以10为单位向上取整 可授予额度=R 1、初始评估包括指入网当时，以及入网后首次动态评估前的这段时期 2、首次动态评估前调整套餐，则信用额度按照以上初始入网信用额度计算规则调整。 3、可授予固定信用额度在系统中记录为固定信用额度的“上限参考值”	0-可授予额度
C级	零	零

图 5.8 初始入网固定信用额度

信用等级	动态评估固定信用额度	
	可授予额度	用户自调范围
A级	无限	0-无限
B级	W=max(前6个月月均消费额套餐月费)*2.5 R= w以10为单位向上取整 可授予额度=max(R当前可授予固定信用额度) 1、max(a,b)是指取a，b中的大值。 2、可授予固定额度在系统中记录为“固定信用额度上限参考值”	0-可授予额度
C级	零	零

图 5.9 动态评估固定信用额度

用户在办理入网服务合同时，业务受理人员告知用户，用户可以根据自身情况选择不同的信用等级，选择不同的等级需要提供相应信用证明文件资料，在网期间如果记录良好，并且在网时间越长，在评估后可以提升信用等级。

本文的突出点在于构造了实时的动态信用评估，把综合了基础额度的用户信用等级评估与临时调整值的消费异动价值评估综合考虑。对现实情况更有应用价值，可以有效

减少因为用户的偶然消费情况而造成停机，给用户带来不便。

5.2 小结

通过三个子模型来构成动态信用价值模型，首先对高危用户进行分离，国内包括金融、电信行业等进行额信用评估研究绝大多数是将所有的用户包括高危用户都考虑在内，这样的结果就是高危用户利用信息不对称也获得正常用户被授予的信用额度，以使公司蒙受损失，本文是首先根据每个用户的历史消费行为对预测用户的将来行为，通过现有的 1 至 6 月份的数据来预测 7 月份的未出账并且确实离网的用户为哪些用户，通过 SPSS Modeler 建模可以得到这部分用户的规则集，约为 65 个集合，把该部分的集合应用到数据库中，对用户根据属性变量进行筛选来对即将离网用户进行分离，不再对该类用户纳入授信考虑范围，以减少该类用户的恶意消费带来的损失。第二部分的用户信用等级评估是对每个用户根据得分享有与之匹配的信用额度的倍数，每个用户的得分越高，则表明越是优质用户，应该享有更高的信用额度，因此得分越高的用户享有的信用额度与之套餐的倍数就越高，是符合现实情况的。此外，在动态信用模型的第三阶段中引入了风险因子，即把用户的多种属性及所选套餐根据重要程度赋予权重来直接作用于每个用户最终的信用额度，在网时间长、所选套餐量多、月消费额度高的用户相比在网时间短、消费额度低的用户将享有更高的信用额度，优质用户不会因为偶然的一次某月的主叫通话时长超出以往平均时长很多时间而造成停机。消费异动会把这种的异常通话带来的偶然的消费异常的信用额度增加给用户，让每个用户放心的消费，主叫通话消费异动与新办业务的消费异动均纳入模型的考虑范围来使得模型更加人性化，符合现实的用户需求。

第六章 总结与展望

6.1 论文总结

本文旨在建立一种动态信用模型，此模型可以判断用户是否是正常不离网用户，并按照用户的属性和消费习惯确定用户的信用等级，还可以根据用户的历史行为或者套餐的变更等偶然的异常消费来对用户赋予实时的信用额度，以减少不必要的停开机给用户带来不好的消费体验和联通某省分公司的损失。

本论文的主要工作包括：

对已有的高危离网用户预测模型及信用评价模型作对比分析，分别研究所用到的数据挖掘算法及其应用的效果。

针对联通某省分公司所面临的问题做实证性的分析，在已有方法的基础上，寻找最优的方法，以解决联通某省分公司所面临的问题。

详细介绍动态信控模型的组成，包括三个部分，分别是离网倾向预测评估、用户信用等级评估和消费异动价值评估。离网倾向模型是基于 C5.0 算法，用户信用等级评估是基于二步聚类算法，每个用户可以根据自己的属性和消费行为得到一定的基础信用额度（高危离网用户的信用额度为 0 ，正常用户按照信用等级来确定），消费异动价值评估是通过把停机比率作为风险因子引入，从而将用户消费的异常变动纳入用户的信用额度考虑范围。系统的实时监控值即为基础信用额度加上临时调整值得到。

将动态信控模型的三个部分与联通某省分公司现有的模型分别进行比较，发现离网倾向预测模型和用户信用等级模型均优于联通某省分公司的现有系统模型，而消费异动价值评估未被现有模型纳入考虑，因此本文所提出的动态信控的基本思想具有很好的应用效果及可行性。

6.2 未来展望

本文的动态信用评估虽然可以有效的减少联通 A 公司的用户的停开机数量及减少损失，但是现有系统有以下一些特点：1.用户信用模型当中的信用等级的划分是基于聚类算法将用户分成不同的类别，因此，模型也就具有了聚类算法所面临的问题，具体在实际中到底聚成多少类更更大的实际价值，也是需要接下来进一步考虑的问题；2.沃家庭用户的信用额度的调整由于现有系统的原因（无法获取数据）没有纳入考虑；3.现有实际情况中可以获得的用户属性有限，比如用户的性别、年龄等属性是无法获取的；4.采用动态信控模型，用户的信用额度可能会经常变动，这部分同样会影响用户感知；5.

有些关键性的指标如套餐内流量，套餐内主叫计费时长等，并且可获得性很大。

未来展望：1.借助社会网络分析方法分析离网用户的行为特征；2.信用等级的划分针对不同的省份采用不同的策略；3.完善现有系统，获取更多的用户属性，在此基础上完善动态信用的评估。

参考文献

[1] 杨力，宋利，候峰. 信用评分的统计模型方法述评 ．统计与决策[J].2006 (7) :141-148

[2] Makowski P. Credit scoring branches out. The Credit world , 1965,(75) :30-37

[3] Davis.R H. Edelman D B.Gammermann A J. Machine Learning Algorithms for Credit Card Applications [J]. IMA Journal of Mathematics Applied in Business and Industry. 1992(4):43-51

[4] Bee Wah Yap，Seng Huat Ong，Fon-YuChiu Using data mining to improve assessment of credit worthiness via credit scoring models［J］.Knowledge-Based Systems 36 (2012) 245–252

[5] Defu Zhang，Xiyue Zhou，Stephen C.H Leung Jiemin Zheng Vertical bagging decision trees model for credit scoring［J］. Expert Systems with Applications 37 (2010) 7838–7843

[6] 姚琦云.多途径提高电信欠费催缴率[J/OL].邮电企业管理，2002（21/25）

[7] 王丽平，李多全基于AHP方法计算电信用户信用度[J]计算机工程与应用. 2008,44(32):232-239

[8] 王娟. 联通某地市分公司用户流失及对策研究[D] 北京邮电大学 10

[9] Keramati, Jafari-Marandi , Ahmadian Improved churn prediction in telecommunication industry using data mining techniques [J] Applied Soft Computing 2014.11.994-1012

[10] Bingquan Huang, Mohand Tahar Kechadi, Brian Buckley Customer churn prediction in telecommunications [J] Expert Systems with Applications 2012.1, 1414-1425

[11] 余文建,沈益昌,杜洋基于Logistic模型的个人信用评分体系研究[J] 海南金融 2007.3:82-87

[12] Ling-JingKao，Chih-Chou Chiu，Fon-YuChiu. A Bayesian latent variable model with classification and regression tree approach for behavior and credit scoring［J］.Knowledge-Based Systems 36 (2012) 245–252

[13] LIU Guang-yuan，YUAN Sen-miao，DONG Li-yan．Prediction of Churn of customers with Data Mining method．ComputerEngineering and Appfications，2007，43(9)：154-156．

[14] Anil Gupta Credit Scoring and Models：Decision Trees [J] PeerCube 2014.10

[15] Hykin, S. (1999). Neural networks: A comprehensive foundation (2nd ed.). Prentice Hall International, Inc..

[16] Eliana Angelini，Giacomo di Tollo，Andrea Roli A neural network approach for credit risk evaluation［J］. The Quarterly Review of Economics and Finance48 (2008) 733–755

[17] Tam, K., & Kiang, M. (1992). Managerial applications of neural networks: The case of bank failure predictions. ManagementScience, 38.

[18] Kryzanowsky, L., Galler, M., & Wright, D. (1993). Using Artificial Neural network to pick stocks. Financial AnalystsJournal, 17

[19] Gang Dong，Kin keung Lai，Jerome Yen Credit scorecard based on logistic regression with random coefficients [J] Procedia Computer Science 1 (2012) 2463–2468

[20] 刘铮铮.基于层次分析法的商业银行信用评级模型研究[D] 2006

[21] 左子叶，朱扬勇.基于数据挖掘聚类计算的信用评分评级 [J] 计算机应用于软件 2004.4 :1-3

[22] Guangli Nie,Wei Rowe, Lingling Zhang, Yingjie Tian, Credit card churn forecasting by logistic

regression and decision tree [J] Expert Systems with Applications 2011.12. 15273-15285

[23] 邓雪，李家铭，曾浩健，等. 层次分析法权重计算方法分析及其应用研究 [J] 数学的实践与认识 2012.4 ： 93-100

[24] Clement Kirui, LiHong, Wilson Cherulyot, Hillary Kirui Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining [J] International Journal of Computer Science Issues 2013.3 165-172

[25] 栾丽华,吉根林决策树分类技术研究 [J]. 计算机工程, 2004,5 : 94-95

[26] 马秀红，宋建社，高晟飞.数据挖掘中决策树的探讨 [J] 计算机工程与应用，2004,1 ： 185-186

[27] 王静红，王熙照，邵艳华，等.决策树算法的研究及优化 [J] 微机发展，2004,9 ： 30-32

[28] 庞素琳，巩吉璋 C5.0分类算法及在银行个人信用评级中的应用 [J] 系统工程理论与实践，2009,12 ： 94-104

[29] 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 :103-105

[30] 王磊，郑任儿.决策树算法的比较研究 [J] 科技信息 2012,30 ： 156-158

[31] MacQueen,J.Some Methods for Classification and Analysis of Multivariate Observations, the 5th Berkley Symposium on Mathematics. Statistics and Probability , 1967,1(1)

[32] 梁礼明，翁发禄，丁元春.神经网络在客户流失模型中的应用研究 [J] 商业研究 2007.2 ： 55-58

致　谢

研究生阶段的学习生涯时间美好而短暂，在北京大学软件与微电子学院三年的求学时间眨眼而过，无比的留恋这段美好的时光。

拙文在选题、写作的不同阶段受到我的导师李杰老师悉心指导，当我在选题不定、文献综述如何展开等论文写作过程中存在困惑时，李杰老师为我指明方向，为我指点迷津，把老师多了解的知识教授于我，帮助我开拓研究思路，精心点拨。李杰老师做研究一丝不苟的态度，踏踏实实的研究精神，对学生们的关爱引导，不仅让我知道自己该如何学习，而且教我明理做人，更是对老师的胸怀佩服不已，虽然只有短短的三年时间，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的。

正是李老师的培养和指导，使我研究问题的思路不断拓展、研究问题的方法更科学，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和家人，是他们一直在支持我，关心我、鼓励我，陪我经历苦难，陪我度过低谷困难。父亲的从小的严格要求让我对生活有了更深的感悟，是我今后人生的宝贵精神财富，在此表达我对他们的深深敬意、感激之情，我爱你们。

我还要感谢北京大学智能感知实验室的兄弟姐妹们。数载相处中，让我也获得了学校生涯最宝贵的友情，大家一起熬夜做小组作业的争执、探讨的场景会让我一直铭记，很多时候聚在一起对生活、理想及人生等诸多问题畅所欲言，尽情的交流，在这种的交流中让我获益匪浅。

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：　年　月　日