分类号_____　　密级_____
U D C _____　　编号　10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕士学位论文

（专业学位）

论文题目　基于属性加权的聚类算法在银行客户细分中的应用研究

研　究　生　姓　名：袁 慎

指导教师姓名、职称：韩君　教授

学　科、专　业　名　称：统计学　应用统计硕士

研　　究　　方　　向：市场研究

提　　交　　日　　期：2020 年 6 月 8 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：袁慎 签字日期：2020.6.8

导师签名：[illegible] 签字日期：2020.6.8

导师(校外)签名：陈波 签字日期：2020.6.8

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：袁慎 签字日期：2020.6.8

导师签名：[illegible] 签字日期：2020.6.8

导师(校外)签名：陈波 签字日期：2020.6.8

Application of Clustering Algorithm Based on Attribute Weighting in Bank Customer Segmentation

Candidate: Yuan Shen

Supervisor: Han Jun

摘 要

在如今互联网金融时代的潮流下，随着国内银行业务范围的扩大、客户量的增长、时间的累积以及数据收集和存储技术的迅速发展，产生了一种“客户数据丰富，但知识贫乏”的现象。银行业的激烈竞争实质上是客户资源的竞争，如何挖掘庞大而多维的数据背后隐藏的潜在市场，如何发现客户的消费需求倾向，如何筛选并挽留易流失的客户等等问题，迫切需要一种能够高效、多维度、精准化的客户细分模型为银行实现企业利益最大化提供决策指导。

聚类算法是客户细分中运用最为广泛的方法，然而传统的 K-Means 算法在实际应用中把所有属性特征按同等贡献度看待，没有考虑不同属性特征对聚类结果可能造成的不同影响，忽略了业务含义。为解决 K-Means 算法所导致的聚类偏差并提升聚类效果，本文在 K-Means 算法的基础上进行改进，通过 Logistic 逐步回归加权的方式筛选重要属性并赋予属性权重，使之能够按属性贡献度对数据对象进行差异化度量，从而设计一种基于属性加权的聚类算法应用到银行客户细分场景中。

本文使用的是从某银行数据库和 CRM 系统中随机抽样的客户全年交易记录及相关信息数据，通过客户的当月 AUM 月日均（金融总资产）这一指标把客户分为低端客户、中端客户和高端客户三组，以为银行带来收益为主要研究目标，从客户基本属性信息、客户标识信息、客户价值信息、RFM 信息、客户交易及动账最值信息五个维度实现银行客户细分，主要分为三个阶段：

第一阶段，三组客户分别运用基本统计分析、趋势分析、业务分析、相关性分析等方法进行变量的选择与确定，以客户 AUM 资产达标为目标变量，应用 logistic 逐步回归模型尝试、比较及业务解读，并通过 ROC 曲线和 Lift 提升曲线的评估验证，最后得到具有可解释性、可靠的相关变量和模型系数。

第二阶段，根据第一阶段所得的相关变量和模型系数使用回归权重设计的方法确定属性加权聚类算法的权重，然后应用传统 K-Means 算法和改进的属性加权聚类算法分别对三组客户依次进行聚类，通过两种聚类算法的可视化结果展示与比较，以及聚类算法性能对比和分离度、紧密度、CH 指数和轮廓系数等有效性评价标准的评估与验证，最终证明属性加权聚类算法的优越性。

第三阶段，应用基于属性加权聚类的客户细分算法，最终将银行客户细分成

13 个小类，对于细分结果进行客户价值分析，合理的判断出需要重点维护的高价值客户类别，需要挽留的易流失客户类别，需要重点发展的潜力客户类别和低价值可放弃的客户类别等等，并提出银行企业维护、发展客户和优化资源配置提供建议。

关键词：客户细分 K-Means 聚类 Logistic 逐步回归 权重设计 属性加权聚类

Abstract

In the current trend of the Internet finance era, with the expansion of the domestic banking business, the growth in the number of customers, the accumulation of time, and the rapid development of data collection and storage technologies, a kind of "rich customer data but poor knowledge" has emerged. phenomenon. The fierce competition in the banking industry is essentially the competition of customer resources. How to tap the potential market hidden behind huge and multi-dimensional data, how to find the customer's consumption tendency, how to screen and retain customers who are easily lost, etc., urgently needs a kind of efficiently, multidimensional and accurate customer segmentation model provides guidance for Banks to maximize corporate interests.

The clustering algorithm is the most widely used method in customer segmentation. However, the traditional K-Means algorithm treats all attribute features as equal contributions in practical applications, without considering the different effects that different attribute features may have on the clustering results. Ignore business implications. In order to solve the clustering bias caused by the K-Means algorithm and improve the clustering effect, this paper improves on the basis of the K-Means algorithm. The important attributes are filtered and weighted by logistic stepwise regression weighting, so that they can be attributed

according to attributes. The contribution degree measures the data objects differently, and a clustering algorithm based on attribute weighting is designed to be applied to the bank customer segmentation scenario.

This article uses a customer ’s annual transaction records and related information data randomly sampled from a bank database and CRM system. The customer is divided into low-end customers, medium-end customers through the indicator of the customer ’s AUM (financial total assets) for the current month. There are three groups of end-customers and high-end customers. The main research goal is to bring benefits to the bank. From the five dimensions of customer basic attribute information, customer contract information, customer value information, RFM information, customer transactions and account value information, the bank's customer details are realized. There are three main stages:

In the first stage, the three groups of customers used basic statistical analysis, trend analysis, business analysis, correlation analysis and other methods to select and determine variables. The customer's AUM asset was used as the target variable. The logistic stepwise regression model was used to try, compare, and conduct business. Interpret and verify through the evaluation of the ROC curve and Lift lifting curve, and finally obtain interpretable and reliable related variables and model coefficients.

In the second stage, the weights of the attribute-weighted clustering

algorithm were determined using the regression weight design method based on the relevant variables and model coefficients obtained in the first stage, and then the traditional K-Means algorithm and the improved attribute-weighted clustering algorithm were applied to three groups of customers. Clustering is performed in turn, and the visualization results of the two clustering algorithms are displayed and compared, and the performance comparison of clustering algorithms and the evaluation and verification of effectiveness evaluation criteria such as separation, compactness, CH index and contour coefficient are finally proved. The superiority of the clustering algorithm.

In the third stage, the customer segmentation algorithm based on attribute weighted clustering was applied, and finally bank customers were subdivided into 13 sub-categories. The customer value analysis was performed on the segmentation results to reasonably determine the high-value customer categories that need to be maintained. The types of customers that need to be retained, the types of potential customers that need to be developed and the types of low-value abandonable customers, etc., and provide advice for banks to maintain, develop customers, and optimize resource allocation.

Keywords： Customer segmentation； K-Means clustering； Logistic stepwise regression； Weight design； Attribute weighted clustering

目 录

1 绪 论

本章首先依据银行数字化转型的时代背景，详细阐述了银行客户细分的需求、目的和意义，引出本文需要研究和论证的主题；然后，从客户细分理论与方法、聚类算法理论和聚类算法在客户细分中的应用三个角度论述国内外的研究现状，并指出本文所要研究和改进的方向，为下文的聚类算法改进和客户细分实证奠定理论基础和实践基础；其次，详细设计了银行客户细分步骤和研究思路框架，并对本文的客户细分实证方面给与综合的概括；最后，突出说明本文研究的创新之处，总结介绍了在聚类算法上的创新和银行客户细分实际应用中的创新，同时指出了本文的一些不足，并提出对于完善不足之处的展望。

1.1 研究背景和研究意义

1.1.1 研究背景

20 世纪初意大利经济学家巴莱多提出了“二八定律”，被各家银行视作金融决策的重要依据，“银行企业 80%的利润来自 20%的优质客户，而其余 20%的利润来自 80%的一般客户”。银行企业相互间的竞争日益加剧，20%的优质客户成为各银行争夺的焦点，按企业成本利润率最大化原则，差异化服务成为银行客户运营策略的主要选择，哪家银行能够运营好优质客户，那么哪家银行就会在市场竞争中占据绝对优势。银行企业的群体庞大而资源却有限，当只能投入有限资源的时候，往往会选择能够带来更高收益的群体，因此，一个能够对银行客户进行精准有效分群的客户细分模型成为银行决策者的迫切选择。

随着互联网金融时代的蓬勃发展，银行业务范围的扩大、客户量的增长、时间的累积、数据收集和存储技术的迅速发展，使得银行企业积累了海量的客户数据，并持续呈现指数级的增长趋势。看似杂乱的数据背后却隐藏着巨大的“宝藏”，若不能及时有效的挖掘数据信息价值，将会导致大量客户的流失与利润的减少，甚至被淘汰出银行业。然而，针对这种“客户数据丰富，但知识贫乏”的现象，传统的数据分析方法无法合理、快速地回答如何通过客户过往的交易记录来发现有价值的信息这一问题。在这种背景下，基于大数据技术的客户细分算法应运而生，尤其是被用来解决大规模数据客户细分案例的效果凸显。

客户细分技术在银行业中得到越来越多的关注，该技术的研究方法也多种多样，常见的构建客户细分模型的方法如人工分群、回归拟合、聚类分析、关联规则和神经网络分析等，然而客户数据的复杂性与多样化间接导致现有方法的不适用性，不管是方法效率还是客户细分精确度上都有待提高。那么在如今海量银行客户数据的基础上，银行企业如何充分利用这些数据的价值，如何为客户提供专项的推荐及服务，这就需要对客户细分模型进行针对性的研究。但并不是所有的客户细分方式都是有效的，这其中主要原因是银行并不能十分有效地将信息技术融合到客户精准分类中，以及应用数据挖掘客户细分技术人员对银行具体业务的认知较为缺乏。因此，如何利用客户细分模型将银行客户的隐藏信息与银行盈利目标相结合，做好客户的精准分类，有针对性地营销产品给客户，才是每家银行能否增加竞争力的关键。

1.1.2 研究意义

在互联网金融的大数据时代背景下，银行业的竞争很大程度上就是数据资源的竞争，客户细分技术无论在理论研究领域、模型构建体系还是实践应用场景中都有着不可替代的作用。如今银行企业的关注点，一是资产概念，二是客户概念，掌控了资产和客户便等同于把握着财富和机遇。本文基于企业成本利润率最大化原则，通过银行客户资产规模（AUM）对研究客群进行分层，在进行客户资产分层之后的基础上应用客户细分模型，更具实际意义。即便低资产客户中也有为银行带来高收益的客户，即便高资产客户中也有带来低收益的客户，对客户进行合理有效的细分，针对性的营销服务，不仅可以优化银行内部资源配置和加强客户关系管理能力，而且可以吸引和保持高价值客户、降低信贷风险并提升银行的市场竞争力。

由于客户群体分化，大众化营销渐显劣势，精准化营销的优势逐渐被发掘，为了更加深层次了解客户，便需要对客户进行精准细分。依据精准细分的结果，可以帮助银行更加清晰的了解各类型客户的消费倾向与行为特征，从而真正获得一群消费能力强、更高收益预期的优质客户群体，有利于银行企业针对客户特定需求进行差异化运营、节约成本、提高效率，最终实现企业真正的盈利。构建客户细分模型，银行企业可以实现海量不同客户群体的精准分群，客户个人可以得

到自己倾向的专项服务，实现银行企业与个人利益的“双赢”。因此，选择恰当的客户细分指标以及客户细分模型方法，对于银行企业实现业务目标显得十分重要。

为了更加全面地剖析庞大而多维的银行客户数据，依据改进的加权聚类算法结合业务经验对海量客户数据建立 Logistic 逐步回归模型筛选指标和设定权重，并依据改进的聚类算法构建基于属性加权的客户细分算法模型。同时针对分好的各类群体，再从基本特征、价值分析、以及消费需求倾向等多维度洞察分群结果特征，进而总结出诸多数据分析结论，制定一套可操作、可落地的银行客户细分策略，为银行客户经营提供指导及落地意义。

1.2 文献综述

1.2.1 客户细分理论与方法研究综述

国际上关于客户细分理论的研究起步比较早，最早起源于美国市场学家 Smith[5]在 1958 年便提出的市场细分的概念，通过把产品差异化与市场细分相结合作为营销策略的制定基础，依据客户群体间的价值、需求和偏好等特征差异性，为客户提供有针对性的产品、服务和营销模式，使得产品属性和营销活动更好满足客户的需要。该理论提出后受到了社会各界的广泛关注，并且随着经济的不断发展，“以客户为中心”的理念逐渐代替“以产品为中心”的理念，从而推动了客户细分理论形成，众多学者也在其基础上进行了创造性的研究与应用。例如 Lazer[6]（1963）以客户的生活方式为研究对象对客户细分，在人口学变量的基础上，通过消费者的生活态度、兴趣和偏好等对企业客户进行细分，刻画各个类别消费者的全貌。如 Haley[7]（1963）从利益细分的角度进行客户细分，他利用具有因果关系的因素而不是描述性的因素来识别市场，能够透过客户表象的行为、态度和动机来挖掘背后的真正利益，为后来学者对客户细分的研究提供了新的方向。又比如 Hughes[8]（1994）提出了 RFM（时间、频度、金额）的概念进行客户细分，即通过最近一次消费（Recency）、消费频率（Frequency）和消费金额（Monetary）三个维度建立 RFM 模型来反应客户的行为特征。但由于传统客户细分方法受到了诸如神经网络之类的新技术所带来的创新概念方法的挑战，逐渐

显出客户细分效率与精准性的不足，例如 Marcus[9]（1998）在 RFM 模型的基础改进，提出了客户价值矩阵模型，不仅能够保留 RFM 模型的简单且低成本的优势，而且优化了客户细分方案，比传统方法更加精准。Albrecht 和 Mario[10]（2001）提出了市场细分和竞争结构的相关理论，认为客户细分的理论不能仅仅关注客户，如果在客户细分过程中考虑竞争对手的信息，则细分结果可以得到很大改善。

国内学者对客户细分理论的研究和实践也有了一定的成就，目前学术界和企业界较为认可的客户细分理论是基于客户价值生涯周期利润（CLP）的客户价值细分理论，陈明亮[11]（2001）认为能否有效地保持有价值的客户是企业成功的关键，依据不同的客户价值确定不同的资源配置方案是企业保持商务战略的首要任务。张国方和金国栋[12]（2003）以客户关系管理为理论依据，把客户价值作为理论核心，提出并研究了客户相对价值细分理论、客户价值生涯周期形态细分理论及其应用策略，解决了传统客户细分理论的缺陷，完善了客户关系管理的理论成果。王扶东和马玉芳[13]（2011）同样认为客户细分是客户关系管理的重中之重，两人全面考虑了客户生命周期价值，从数据挖掘技术和群体决策技术角度出发，用实际数据证明了客户细分理论的实践意义。徐翔斌和王佳强等人[14]（2012）运用 Hughes 提出的 RFM（时间、频度、金额）的理念，改进了 RFM 模型的原有方式，引入总利润属性进行比较分析，为电子商务客户细分的应用实践提供了相应理论策略。廉亦璇[15]（2013）综合比较了我国城市商业银行和国外商业银行在客户细分理论方面的应用策略，提出了城市商业银行在个人理财业务领域如何进行客户细分的建议。许荻迪[16]（2015）也在大量客户细分理论研究文献的基础上，立足于客户细分理论的发展现状与趋势，从银行业的实际应用角度丰富并完善了该理论在银行商业实践的应用。

从国内外关于客户细分理论的研究来看，在不同的应用场景下已经形成了一套独特的体系，也有着多种不同的客户细分方法。比较具有代表性的是刘英姿和吴昊[17]（2006）关于客户细分方法的研究综述，他们从维度和细分技术两个角度将客户细分方法大致分为四类：一是人口统计细分，如 Chou 等人[18]（2000）使用客户的人口统计信息从大量候选人中识别潜在客户，运用数据挖掘技术寻找有可能成为潜在客户倾向的人口统计信息；二是生活方式细分，如 Wells 等人[19]（1971）提出的用 AIO（activity、interest、opinion）来代表生活方式的概念，引

发了后来学者对生活方式细分方法的研究；三是行为细分，如林盛[20]（2006）、蔡玖琳[21]（2015）和施荣晗[22]（2018）等人运用 Hughes 提出的 RFM 模型理念，分别在电信客户、零售业客户和商业银行客户领域对行为细分方法进行了研究；四是利益细分，如郑琦[23]（2000）认为市场细分理论对企业市场营销活动起主要指导作用，通过利益细分调研结果论述其观点的正确性。随着时代的发展，这四类细分方法的研究对于问题的研究越来越表现出不足，因此也引发了众多学者对客户细分方法改进的潮流，比如綦欣德[24]（2013）摆脱传统的单一方法细分，由静态客户细分模型的研究转向动态变化过程研究，从多维度、动态性与预测性三个新视角探讨客户细分方法。曾小青和徐秦等人[25]（2013）基于数据挖掘理论，提出了提出一种过程完整的客户细分新方法，利用销售数据表明方法的有效性。王颖晖[26]（2009）指出客户细分不再局限于客户行为特征变量，提出了基于态度变量的客户市场事后细分策略，运用 K-means 聚类分析、SOFM 和支持向量机方法对比实施客户细分方案。

1.2.2 聚类算法理论研究综述

现如今，客户细分理论的发展和研究已经较为完善，但是一劳永逸的客户细分方法无法通用于复杂且多变的客户场景，采用不同的客户细分方法分析同一客群，其结果往往迥然相异。随着金融经济的发展，衍生出了适用于不同场景的大数据客户细分方法[27]，其中最主要的方法便是聚类算法。

在聚类算法的发展历史上，其更新方向有两个：一是改进现有的聚类算法：针对 K-Means 聚类存在四个缺陷，对于离群点和孤立点敏感；k 值选择；初始聚类中心的选择；只能发现球状簇，即改进距离，也是本文所改进之处。二是发明新的聚类算法：如最先普及的 K-Means 算法，后来推出 K-modes 算法，再推广到 K-prototypes 算法，甚至基于深度学习的神经网络聚类算法等。早在 20 世纪 60 年代，MacQueen[28]便首次提出了 K-Means 算法，简单易用的性质使其仍然活跃在各种聚类场景中，但是 K-Means 算法仅适用于数据集属性连续的情况，对于离散属性数据集的处理则存在很大的误差。为解决 K-Means 算法不能处理离散数据的局限，Huang 等人[29]（1998）提出了 K 众数（K-Modes）算法，采用属性差异度来代替 k-means 中的欧氏距离，用以解决离散属性的数据聚类。20 世

纪尾声，Huang 等人[30]又进一步将 K-modes 算法推广到混合属性数据层面，提出 K 原型（K-prototypes）算法。针对处理诸如银行客户复杂行为的数据，很多学者也对聚类算法进行了改进。例如 Zhi 和 Gong[31]（2010）通过引入进化算法框架提出的新型无监督 k-原型聚类算法（EKP）在合成和真实数据集上更稳健并且产生比 k-prototype 算法更好的结果。Hsu 等人[32]（2007）提出了基于方差和熵的聚类算法 CAVE，它能够挖掘客户数据以进行客户细分和目录营销的应用。后来，Hsu 等人[33]（2008）又提出了一个增量聚类算法，有助于表达分类值之间的相似性，并且还统一了数值和分类值的距离测量。Kim 等人[34]（2014）使用 MapReduce 为大数据提供有效的基于密度的聚类算法，把基于密度的聚类算法与 MapReduce 框架很好地扩展，证实 DBCURE-MR 可以有效地找到簇，而不会对具有不同密度的簇敏感。

相对于国外对聚类算法的研究，国内学者也对原有的聚类算法进行改进。例如，张晓峰[35]（2010）在传统的 K-means 聚类算法基础上为数据属性赋予权值，引入特征权重并重新构造了针对不确定数据集的聚类算法从而提高聚类精度。陈韡和王雷等人[36]（2010）改进 K-prototypes 算法处理混合属性数据时计算分类属性相异度的公式，使其较改进之前的算法具有更好的稳定性和更高的精度。熊平和顾霄[37]（2014）基于 K-Means 聚类在计算样本与质心的距离时为各属性赋予相应的权重，提出了一种应用拉格朗日乘数法自动计算最优的属性权重的聚类算法。赵兴旺和梁吉业[38]（2016）在信息熵的概念下，提出了一种针对数值型和分类型属性的混合数据属性的加权聚类算法，证明能够有效的解决高维混合数据聚类中属性加权问题。黄晓辉和王成等[39]（2019）基于简单高效的 K-Means 聚类算法，通过设定目标函数优化求解得到算法参数的更新迭代公式，来设计一种集成簇内和簇间距离的加权 K-Means 方法。

1.2.3 聚类算法在客户细分中的应用综述

国内外对于聚类算法的研究已经较为全面，聚类算法作为一种非常重要的数据挖掘技术，在客户细分领域的研究占据重要地位，并且国内外针对聚类算法在客户细分领域的研究与实践已经取得了相当不错的成就，涉及的行业也比较广泛，尤其在银行业的应用场景下已经形成了一套独特的体系。例如，Zakrzewska 和

Murlewski[40]（2005）利用庞大且具有多维性的银行客户数据，比较 DBSCAN 聚类、K-Means 聚类和基于两阶段聚类三种方法进行客户细分的效果，得到能够应用到银行客户细分中最有效和可扩展的算法。孙晓霞[41]（2006）详细探讨了聚类分析技术在客户细分领域的应用，将模糊 c 均值（FCM）聚类算法应用到银行客户细分实验中，又在 FCM 算法的基础上去除空簇，提高了算法效率。花海洋和赵怀慈[42]（2008）在银行个人金融领域实现客户细分，将 DBSCAN、K-means 和 X-means 三种聚类算法对比分析，提出最适合银行业客户细分的算法，建立了一套执行效率高、可扩展性大和异常点检测能力强的银行客户细分模型。樊宁[43]（2011）从传统聚类算法对初始聚类中心敏感的角度出发，对 K-Means 进行改进优化，克服了 K 均值算法易陷入局部最优值的问题，提出一种基于改进的 K 均值聚类的银行客户细分方法，用以提高客户细分准确率，为银行决策者提供有效的参考建议。瞿小宁[44]（2011）在对商业银行客户细分时，解决了传统 K 均值确定初始聚类中心问题的缺陷，提出了一种基于粒子群优化 K 均值聚类的商业银行客户细类模型，从客户细分聚类方法上提高算法的收敛速度，增加细分准确率。秦秀洁[45]（2014）和樊仙仙[46]（2016）均基于 CRM 客户关系管理的概念角度出发，针对传统聚类算法的优缺点取长补短，提出改进的客户细分算法和实施方法，并将其应用于银行客户关系管理中，为银行的经营决策提供有力支撑。李涛[47]也在 2016 年第二届今日财富论坛会议中也表明了对商业银行客户细分中的观点，认为商业银行领域的竞争日益激烈，商业银行若想要获得优势地位，就必须重视吸引客户的能力，必须重视客户细分。于化龙和韩雪峰[48]（2018）从确定最佳聚类数的角度对客户细分聚类算法进行改进，通过定义类间最大相似度均值（AMS）确定初始值中心来实现银行客户的细分，提高银行客户细分正确率，使银行的收益最大化。

基于聚类算法的客户细分技术不仅仅在银行领域拥有重要地位，早在 19 世纪就已经被广泛地应用于其他许多领域，其应用场景也变得错综复杂，例如零售、金融股票、对外贸易、能源、电子商务、电信、旅游航空等行业领域。例如在零售行业，Bonoe 和 Rohem[49]（2002）运用基于 Hopfield-Kagmar 聚类的人工神经网络技术进行客户分群，并证明较之 K-Means 算法具有更好的优越性。在金融股票行业，Shin 和 Soho[50]（2004）应用模糊 K-Means 聚类算法对股票客户细分，

对比传统 K-Means 和 SOM 聚类的结果，证明模糊 K-Means 聚类算法更稳健。在对外贸易行业，Golsefid 和 Ghazanfari 等人[51]（2007）通过寻找聚类的最佳数量，应用 RFM 模型进行客户细分研究并分析每个群集的相对盈利能力。Jeff 和 Ramandeep[52]（2019）在能源电力领域将 k 均值聚类应用于居民用电消费者进行客户细分，研究各类别居民用电消费的影响因素，以期望节约电网成本。陈汉思和张磊等人[53]（2019）基于客户使用模式之间的距离对对智能手机客户进行细分，从客户运营数据中研究客户间的异质性，寻求客户细分的重要因素。在电子商务领域，李鑫鑫[54]（2012）对传统客户细分模型进行改进，提出了基于半监督近邻传播的改进 k-means 算法，在某网站客户细分的应用上实现了比较好的性能。电信行业客户与银行客户类似，存在着日益庞大的客户群体，对于电信行业客户细分的研究不逞多让，如陈治平[55]（2007）和武森[56]（2008）等人将实际案例与聚类分析技术相结合，在电信客户细分中都提出了一种解决电信客户细分的应用模型，并用实际数据验证其有效性，以提高电信企业的核心竞争力。在航空客户的研究中，张利利和马艳琴[57]（2019）对航空公司的客户行为特征进行分析，利用 K-均值聚类对客户细分并挖掘出有价值的客户，从而达到提高上座率和效益的目标。在旅游行业，汪永旗和王惠娇[58]（2015）借助 MapReduce 框架对聚类算法进行改进，构建了一种多指标的 RFM 扩展模型，实现了客户细分在旅游行业的应用价值。

1.3 研究思路及框架

银行客户细分的目标是将现有的客户群体按某种规则分成若干子群，使得不同子群之间的客户具有显著的差异特征，同一子群内部客户具有相似的特征，为银行企业能够依据细分群体间的属性差异运营客户，挖掘客户的潜在价值并实现利润最大化提供指导性建议。本文限于传统 K-Means 聚类算法忽略实际业务含义而把数据指标特征同等重要性看待的缺陷，对 K-Means 聚类算法进行优化改进，设计一种基于属性加权的聚类细分算法，借以提升聚类效果，使其能够结合业务含义对庞大而多维的银行客群进行更精确的细分。

本文使用的是从某银行数据仓库和 CRM 系统中随机抽样的客户交易记录及相关信息数据，从客户基本属性信息、客户标识信息、客户价值信息、RFM 信

息、客户交易及动账最值信息五个维度应用客户细分模型，经过对基于属性加权聚类算法的设计、比较、分析、评估和验证，得到可以应用于实际的银行客户细分算法和精准有效的客户细分结果，为银行决策者提供更加准确的指导性建议，从而提升银行的竞争力。本文的研究思路主要分为以下几个步骤：

首先，通过客户的 AUM 资产月日均（金融总资产）这一指标把客户分为低端客户、中端客户和高端客户三组，针对每组客户分别进行数据清洗、数据变换，并运用基本统计分析、趋势分析、业务分析、相关性分析等方法对数据进行分布探索，从而选择和确定属性变量。

然后，三组客户分别以客户 AUM 月日均是否提升达标为目标变量，应用 Logistic 逐步回归模型，通过 ROC 曲线、AUC 指标、Lift 提升曲线和 Lift 提升值进行模型的评估与验证，得到具有可解释性、可靠的相关变量和模型参数。

其次，以三组客户 Logistic 逐步回归模型结果使用回归权重设计的方法分别构建加权聚类算法的属性权重，应用改进的属性加权聚类算法分别对三组客户依次进行聚类，与传统 K-Means 算法聚类结果进行对比，并通过聚类算法的性能度量与分离度、紧密度、CH 指数和轮廓系数的聚类结果有效性评价标准与 K-Means 聚类进行比较，证实改进算法的优越性。

最后，应用模型实现得到的三组客户细分结果进行客户价值分析、行为特征分析以及精准化营销推荐等数据挖掘任务，为银行企业实现利润最大化提供决策建议。

本文研究内容的主要思路框架如图 1.1 所示：

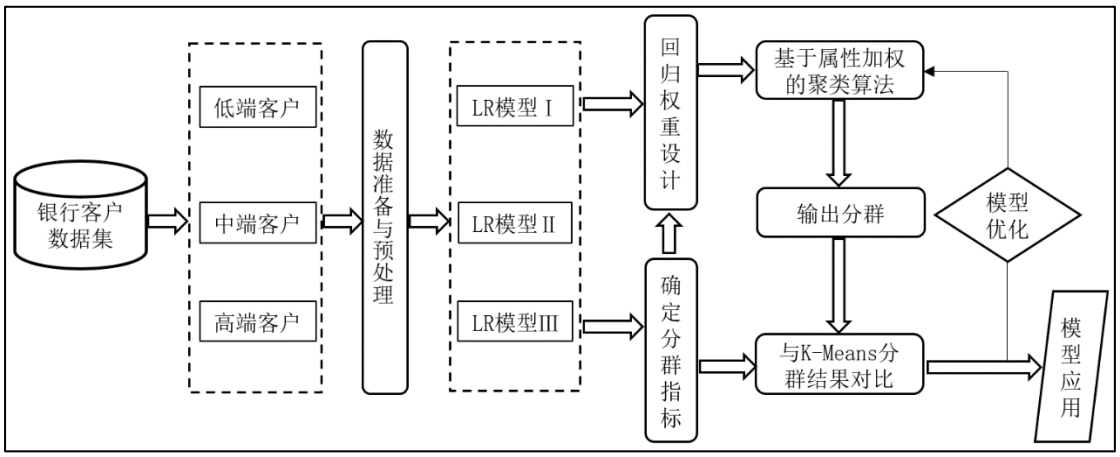


图 1.1 主要思路流程图

1.4 研究创新与不足

1.4.1 研究创新

聚类算法的目的是将数据集人为地划分成若干类，以揭示这些数据分布的真实情况，期望得到的聚类结果能够带来更好的效益。然而，多种多样的聚类算法无非是聚类算法与其他方法的结合、多种聚类算法的集成、针对算法缺陷的改进优化等等，不可否认其功能的强大，但是大多数聚类算法，都存在一定的局限。本文的创新点则在于依托传统 K-Means 算法，针对银行客户数据的特定场景，根据具体的业务含义及需求预先进行模型的规划，改进一种基于属性加权的聚类算法应用到银行客户细分场景中。相对比传统 K-Means 聚类算法忽略实际业务含义而把数据指标特征同等重要性看待的缺陷，经改进的加权聚类算法在其基础上赋予权重，综合考虑业务指标含义，更适用于银行客户数据处理，加权聚类算法更是解决了 K-Means 算法只能识别球状簇的劣势，而且由于其 Logistic 回归加权的特性，能够处理像银行客户数据这样高维度高规模的数据集。

本文提出的基于属性加权聚类的客户细分算法优势有三点：第一，通过多组客户的细分结果与聚类评估综合比较，证实了改进后的聚类算法性能与其它算法比较具有一定的优越性；第二，能够有效的处理高维且数据量庞大的银行客户数据，客户细分结果的客户价值与实际效益得到了全面的分析；第三，本文在数据处理和指标选取上突破了传统客户细分在指标选取上的漏洞，客户数据全部原始形态，包括指标的生成和指标宽表的搭建，以及从一级指标到三级指标层层划分，层层递进地对初始目标数据集进行分析，通过数据清洗、数据变换、基本统计分析、趋势分析、业务分析、相关性分析等多种方法进行相关操作，最终得到与实际业务相符合的模型变量和参数。

1.4.2 不足之处

聚类算法的改进并不是一成不变的，需要结合特定的应用场景选择更加合适的方法，没有任何一种算法能胜任任意形状、任意分布数据的精准聚类，或多或少的有一些局限性。同样的，本文提出的基于 K-Means 改进的加权聚类算法也

有一些不足之处。

（1）加权聚类算法本质上是为庞大而多维的银行客户行为数据所设定的，或者说是为拥有大量客户行为数据的应用场景设定的。该加权聚类算法对于属性特征的筛选有很大的优势，但是由于应用场景的局限性，它高度依赖于正确的数据表示和业务指标含义的深刻理解。并且由于该算法采用的是 Logistic 逐步回归筛选变量与确定权重，因此对于一些非线性问题的解决并不能给出权重设计方案。

（2）加权聚类算法同样的具有 K-Means 聚类算法相一致的一些局限。如算法初始聚类中心的确定问题，一旦初始值选择的不好，可能无法得到有效的聚类结果，并且在更新簇中心时同样采用的均值度量，对于孤立点的衡量效果依旧有一定的误差。

（3）由于银行客户数据的保密性，更多的细分指标和最新的数据获取不到，缺乏未来的实际验证，略显不足，但是可以根据数据库的不断扩充和各种方法的比较使得客户细分精准性更高。

2 基于属性加权聚类的客户细分算法

前文对于本文的研究背景意义、国内外相关文献综述、研究思路框架和本文的创新与不足进行了详细的阐述，本章则是本文的理论核心部分，重点介绍本文所用的基于属性加权聚类的客户细分算法实现过程，包括属性权重设计方案和聚类算法改进两方面内容。第一，属性权重设计方面，采用的是 Logistic 逐步回归权重设计方法，简单介绍了 Logistic 逐步回归原理和熵权法、因子分析法与回归系数法等多种属性权重设计方法，最终确定了本文的 Logistic 逐步回归权重设计方案；第二，聚类算法改进方面，对于多种聚类算法的分类、局限性和相关改进简单介绍，最后确定本文改进的聚类算法和算法实现流程，并且对本文算法的性能和有效性评价方式详细阐述，为后文的银行客户细分实证部分奠定坚实的理论基础。

2.1 Logistic 逐步回归权重设计

2.1.1 Logistic 回归原理

Logistic 回归又称对数几率回归和逻辑回归，是银行业数据建模过程中较为常用的算法，其原理不如随机森林、神经网络和 XGBoost 等算法高深，但在银行业中 80%的预测模型使用的是逻辑回归，其优势是其它算法无可替代的。具体原因有如下三点：一是 Logistic 模型训练速度快、预测准确、模型复杂度低且更加稳健，在金融银行领域中可以更高效的实现业务目标；二是模型直观，变量系数含义易于理解，在如今银行领域的应用环境多为 SAS 以及 SQL，当客群变化导致模型效果下降的时候，Logistic 模型可以迅速的找到原因，更改参数优化模型效果；三是基于模型的结果是概率值，更容易让人信服，例如本文分析问题过程中可以得到每个客户的资产达标概率，既可以实现客户资产的预测，也可以排序确定客户重要程度。综上，本文在此使用逻辑回归算法从大量的银行客户历史数据中找出有哪些关键特征数据，同时给出关键特征的重要性程度，用于权重设计来进行加权聚类算法模型的实现。

Logistic 回归算法不同于线性回归，后者一般是找出因变量与自变量可能存在的线性关系并且确定参数值，而逻辑回归算法利用 Logistic 函数将因变量的取

值范围定义在 0~1 之间，表示因变量为 1 的概率。其原理如下：

本文所用的 Logistic 回归的因变量是二分类的，只有 0 和 1 两种取值。假设因变量取 1 的概率是p，则因变量取 0 的概率则为$1-p$，事件的优势比则为两者概率之比$p/(1-p)$，取值范围为$(0,+\infty)$，优势比越大，事件发生的可能性越大。对优势比取自然对数即为 Logistic 变换$\mathrm{Logit(p)}=\ln(\frac{p}{1-p})$。

令$\mathrm{Logit(p)}=\ln\left(\frac{p}{1-p}\right)=\mathrm{z}$，则$p=1/(1+e^{-z})$即为 Logistic 函数，取值 0-1 之间，其分布如图 2.1 所示：

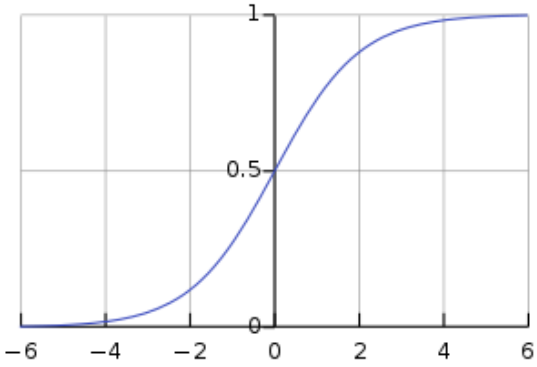


图 2.1 Logistic 函数

如图 2.1 所示的“S”型分布曲线则是 Logistic 函数，当分布概率为 0.5 时的中心点附近增长很快，而在两端增长很慢。因此次在预测时可以很好地把中心点附近的数据轻易分类，并对预测结果可以进行概率排序。

Logistic 回归模型是$\ln(\frac{p}{1-p})$与自变量之间的线性回归模型，如式 2.1 所示：

$$\ln\left(\frac{p}{p-1}\right)=\alpha+\beta_1x_1+\cdots+\beta_px_p+\varepsilon \quad (2.1)$$

或者如式 2.2 所示：

$$p=e^{\alpha+\beta_1x_1+\cdots+\beta_px_p+\varepsilon} \quad (2.2)$$

其中，p为事件发生（因变量为 1）的概率，自变量$x_1,x_2,\cdots,x_p$可在任意范围内取值，α指在自变量$x_1,x_2,\cdots,x_p$全部取 0 时优势比的自然对数，β_i指某自变量x_i每增加一个单位，事件发生的概率增加为原来的e^{β_i}倍。

本文利用 Logistic 回归模型的目的是为了更好地筛选出与业务目标相关的变

量，同时确定变量间的权重，因此模型的有效性检验是不可或缺的。检验方法主要为以下三种：一是划分训练集（80%）用来进行模型构建，和测试集（20%）用来进行模型评估，使得模型具有更好的泛化能力和效果；二是利用 ROC 曲线和 AUC 值作为“测试误差”来作为展现和度量模型好坏的一个标准，即在尽量少的误分类基础上，尽可能多的预测资产实际提升的客户，衡量模型整体预测能力；三是使用 lift 曲线和 lift 提升值来评估使用模型相比不使用模型时的客群响应率，即我们采用这种模型的客群实际提升效果，衡量模型提升效果能力以及模型属性的重要性程度，最终将评估好的模型应用到加权聚类算的的权重设计过程中。

2.1.2 逐步回归原理

在高维度的银行客户数据中，要筛选出能够进行加权聚类算法的变量是很复杂的过程，不仅仅包括业务分析和数据预处理过程的筛选，更重要的是通过模型本身进行选择，本文则是在 SAS 环境中，利用 PROC LOGISTIC 过程步，对可能影响 AUM（金融资产）提升情况的诸多变量中挑选一批关系较好的变量作为影响因子，其中采用的主要方法便是逐步回归。

一般而言，我们一方面要保留对目标变量影响显著的因子，另一方面又要使得 Logistic 回归方程更加拟合，只有这样才能更准确的筛选变量。PROC LOGISTIC 过程步中，对于变量的筛选方式有三种：Forward、Backward 和 Stepwise。Forward 指变量逐个进入模型，每次进入 P-value 最小的变量，直到未进入的变量都不显著；Backward 指所有的变量先一次性进入模型，然后逐个剔除 P-value 最大的变量，直到保留的变量全都显著；Stepwise 指变量逐个进入模型，但是进入新的变量以后，会重新审查所有已进入变量的 P-value，如果进入新的变量导致原来变量的 P-value 从显著变成不显著，则把原来的变量剔除。本文使用的 Logistic 逐步回归方法的基本思路则是运用向前向后逐步筛选法（Stepwise）自动地从大量的可供选择的变量中选取最重要的变量，据以建立 Logistic 回归分析的预测或者解释模型，来达到筛选变量与确定属性权重的目的。

2.1.3 属性权重设计

本文所改进的赋予属性权重的聚类算法，其中关键点是在于属性权重的确定。从理论角度来讲，属性的权重是一个相对的概念，即该属性在整体中的相对重要程度，是根据对被衡量对象的作用轻重进行区别划分。从实践角度来讲，权重和贡献度、变异系数、回归系数、因子载荷系数等关联性很大，很多时候不过是研究方法、计算方法下的不同称呼。从赋权方式上看，目前常见的权重确定方法分为主观赋权法和客观赋权法。

主观赋权法是指依据研究目的和评价指标内在含义的不同，从主观上判断各个指标对整体的贡献程度。常用的主观赋权法如：专家调查法，又称为德尔菲法，是一种通过人为评定得分综合确定的定权方法，其优点是意见集中、评定统一，缺点是人为干涉打分难以保持权重的合理性；优序法，通过对指标两两相对比较，将结果在 n×n 的表格中进行统计，计算各指标的得分总和与总数的比值，则为该指标的权重；层次分析法，是一种多目标多准则的决策方法，该方法基于人们对于每一层次中各因素相对重要性，将评估目标分解成一个多级指标，逐层两两比较排序，从而对每一层中各因素的相对重要性给出判断。此外，还有二项系数法、环比评分法、最小平方法等。

客观赋权法是指从实际数据出发，由数据本身内在关系提取出权重的方法，某一指标的权重可以理解为该指标在数据集中的变异程度和对其它指标影响程度的综合度量。常用的客观赋权法如：熵权法，是根据指标的信息熵，即信源的不确定度大小来确定客观权重，虽是实际中最广泛应用的方法，但是没有考虑指标与指标之间的影响（如相关性、层级关系等），而且该方法对样本的依赖性较大，随着建模样本不断变化，权重会发生一定波动，导致权重可信度较低，因此并不适合银行客户细分的研究；因子分析法，用主成分得分矩阵作为计算基础，对每个指标计算共性因子的累积贡献率来确定对应变量的权重，相对于主成分分析法来说因子间增加了可解释性，但是计算过程中仅通过二级指标计算，没有涉及到整体指标，并且在计算因子得分时，采用最小二乘法可能会失效，因此也不适合银行客户间的加权聚类分析；回归系数法，也是本文主要采取的方法，首先利用回归分析方法确定回归模型，然后以自变量的系数绝对值为计算基准，最终确定各变量对应的权重。其好处是可以根据研究对象及数据分布情况的不同，针

对性的应用不同的回归模型，如线性回归、非线性回归、主成分回归、Logistics回归、岭回归、偏最小二乘回归等，根据最终模型的系数，进行标准化处理计算权重。此外，还有相关系数法、主成分分析法、离差及均方差法、梯度下降法等。

主观赋权法可依据自身知识经验和实际问题合理的确定各属性权重，但其结果主观随意性，缺乏客观评价，同时对决策分析者产生负担，有很大局限性。而客观赋权法综合考虑各指标间的数量关系，从实际数据本身内在关系来确定权重，结果更具有客观性和说服力。综合主观和客观的赋权方法优劣，为减少赋权的主观随意性，保证权重与属性间的相互关系，本文同时考虑指标数据之间的内在规律和研究问题的实际意义进行组合赋权，采用回归系数方法，利用主观决策、数据分布探索、相关性筛选与逐步回归等方法筛选属性之后的 Logistics 回归系数确定权重。回归系数表示的是自变量与因变量之间的关系，对其进行标准化之后，可以很好的成为属性的权重，其公式为$w_i = |b_i| / \sum_{i=1}^{m} |b_i|$，其中$w_i$表示第 i 个属性的权重；$b_i$为 Logistic 逐步回归系数。

2.2 聚类算法的改进

聚类技术属于无监督学习的一种，与分类、回归等监督学习方法不同的是，在聚类中那些表示数据类别的分类或者分组信息是没有的，并不需要使用训练数据进行学习。简单地说，就是在聚类时，我们只需要把相似的个体按照某个特定标准（如距离准则）聚到一起即可，而不用去深究每一类别的含义，只要保证同一个簇内个体间的相似性尽可能大，同时使得非同一簇个体间的差异性也更大，便达到了聚类的目标。聚类既可以是一个单一的研究方法，也可以作为分类、回归等其他学习任务的阶段性过程。例如本文对于银行客户的聚类，根据聚类结果将历史客户每个簇定义为一个类后，可以确定类特征，再基于这些类训练分类模型，用于判别新用户的类型。

2.2.1 聚类算法分类及局限性

目前，聚类算法被广泛应用在客户细分中，根据不同的聚类目的、多样的数据类型以及复杂的应用场景，衍生出多种多样的聚类算法。采用不同的聚类算法分析同一问题得到的结果也存在明显差异，因此要结合实际来选择合适的聚类算

法，这些算法有些仅能聚出簇状的类，有些算法仅适用文本数据，而有些算法具有较强的抗噪能力，还有些算法需要人工干预。若按照划分方式的不同则可细分为以下几种类型：

（1）基于划分的算法（Partition-based methods）

基于划分的算法也被称为基于距离的算法，其原理是对于一个拥有 n 个数据点的集合，使用基于距离的相似性度量方式在数据集上进行一层划分。划分算法通过迭代方式渐进地提高聚类质量，很适合寻找中小规模数据集中的球状簇，其中最为典型的 K-Means 算法，是应用最为广泛的聚类算法，其优势速度快，简单易行，对于比较大的数据集运行效率较高。该算法通过遍历每个数据对象与各聚类中心的欧氏距离，划分到最近聚类中心所代表的簇中，并计算簇内对象的均值作为新的中心点，继续上述过程，迭代至目标函数值最优。此外，还有在 K-Means 算法基础上不断改进的算法，例如 k-modes、k-medians、k-medoids、kernel K-Means、k-prototypes、CLARANS 等算法。

K-Means 算法最为具有代表性的划分算法之一，缺陷也不容忽略，因此以 K-Means 算法为基础的改进与时俱进。K-Means 只用于数值型数据，不适用于分类型数据，所以产生了针对分类型数据的 k-modes 以及针对混合型数据的 k-prototypes；K-Means 对初始值设置敏感的缺陷，间接导致 K-Means++、intelligent K-Means 等算法应运而生；k-medoids 和 k-medians 解决了数据中噪声和离群值的影响；kernel K-Means 更是能够解决非凸数据的聚类。本文则是根据代表性算法 K-Means 为基础，对其增加了属性权重，并对分类型数据进行设定哑变量和连续化处理，更具实际意义。

（2）基于层次的算法（Hierarchical methods）

基于层次的算法是把数据对象以类似层次方式的聚合或分解来实现聚类，这个过程可以是基于距离，也可以是基于密度的聚类，直到某种条件满足为止。具体可以分为两种类型：第一类是自底向上合并数据集的层次聚类；第二类则是自上而下分裂的层次聚类。基于层次的聚类算法具有可解释性强，速度快的特点，其优势在于聚类的个数在层次合并或分裂的过程中自动获取，并不需要提前设定，而且能帮助解决 K-Means 不能解决的非凸数据。但是在层次聚类过程中，由于无法撤销步骤中的合并或分裂操作，尽管可以减少计算代价但同时也具有无法修

正错误操作的缺点，并且整个过程时间复杂度高，且难以处理不同大小的聚类簇以及凸形状。

一般的层次聚类算法由于其局限性仅适用于小数量级的场景，例如对中国省会城市的聚类，因而经改进后的 BIRCH 算法，则解决了这种局限，主要在数据体量很大的场景下使用，采用聚类特征树进行多步骤优化，但却仅用于数值型属性的聚类；改进的 CURE 算法与传统的聚类算法不同在于采用收缩因子使每个点更加紧凑并且靠近类中心，对噪音的敏感度不高，适合聚类非球状的数据集；层次聚类改进的算法中，Chameleon 算法每层划分时细致检查对象相互之间的关系，聚类效果被比 BIRCH 更强，可以处理非常复杂形状的簇，但经改进的算法仍然没有摆脱时间复杂度高、一步错步步错的局限。

（3）基于密度的算法（Density-based methods）

基于密度的算法是依据数据对象分布的密度来实现聚类，通过遍历每个数据对象规定半径以内数据对象的个数作为密度值，将密度相对一致的临近数据点归为一类，只要超过事先设定的密度阈值，则为高密度点，将其加到所获取的类中，否则为边缘点，直到所有数据对象遍历完成为止。该算法适用于更多类型的数据集，其中最具有代表性的基于密度的聚类算法是 DBSCAN。该算法引入“核心对象”和“密度可达”的概念以适用于更多类型的数据集，其特点是能克服了 K-Means 等基于距离的算法缺陷，不需事先确定要形成的簇类的数量而且能发现任意形状的簇，同时也对噪声数据的处理比较好，但是该算法对如密度半径 eps 和密度阈值 minPoints 等参数设置敏感，而且难以确定噪音闭值。针对基于密度聚类算法的局限性，也有许多改进的算法。如 GDBSCAN 算法能够适用于空间数据对象，OPTICS 算法填补了 DBSCAN 的一些缺陷，又比如 FDC 算法能够明显提高聚类效率等等。

（4）基于网格的算法（Grid-based methods）

基于网格的算法是通过将将数据对象集映射到已划分的数据空间网格单元中，并对每个网格单元计数作为密度，依据预设的阈值临近形成一类。该方法最大的优势就是聚类速度快，依赖于网格单元的数目，而不是数据对象的数目，可以处理很大的数据集并且发现任意形状的簇，但是该方法是以牺牲聚类结果的精确性为代价的，因此经常与基于密度的算法结合使用。针对基于网格的算法改进，

往往从参数敏感、精度不高、维度灾和不规则分布数据等角度进行优化，其中WaveCluster（利用小波变换聚类）、CLIQUE 算法和 STING（统计信息网格聚类）是该类方法中的代表性算法。

（5）基于模型的算法（Model-based methods）

基于模型的算法是一种“软聚类”方法，通过假定模型的方式来寻找数据对给定模型的最佳拟合来实现聚类。主要分为两大类，一是以 GMM（高斯混合分布）算法为代表的基于概率模型的算法，这里的概率模型假设数据是根据潜在的概率分布生成的，属于概率生成模型，每个属性上的概率分布是彼此独立的，然而往往很多特征是具有相关关系的。二是以 SOM（自组织特征映射网络）算法为代表的基于神经网络模型的算法，该算法通过自动寻找数据对象中的本质属性和内在规律，自组织地改变网络结构与参数，实现从输入层到输出层的降维映射。其优势在于最后的样本点以概率的形式聚到每一类中，而且每一类的特征也可以用参数来进行表达；其局限性在于执行效率不高，且不适用于无法建立模型的小规模数据集。

（6）其他聚类算法

目前，还有一些其他新的聚类算法被提出，如基于模糊的聚类算法、基于约束、谱聚类、核聚类、量子聚类等等。基于模糊的聚类算法（FCM 模糊聚类）克服了非此即彼的分类缺点，以模糊集合论为数学基础进行聚类分析，致力于获得数据集中数据对象划分到不同类的不确定程度，对于满足正态分布的数据聚类效果会很好；基于约束的聚类算法用两点之间的障碍距离取代了一般的欧氏距离，能很好地利用以及表达约束条件；谱聚类算法是将聚类问题转化为图的最优划分问题，是当前聚类领域研究的热点；核聚类算法是利用核函数把输入的数据样本映射到高维特征空间实现聚类，与 SVM 算法思想异曲同工；量子聚类算法是用物理学量子理论解决一些传统聚类算法无法解决的几种聚类问题，为改进聚类方法拓宽了领域。

2.2.2 差异化度量方式改进

聚类算法简单地说就是一种确定样本之间距离的过程，而距离的精确度也间接的衡量了聚类算法的精确度。一般而言，K-Means 聚类算法通常采用欧氏距离

作为相似性度量方式，而本文所改进的聚类算法对数据对象赋予属性权重，使之能够按属性贡献度对数据对象进行差异化度量，在 K-Means 框架下更加客观地度量数据中对象与类之间的相异性。该差异化度量方法的相对优势是能够结合业务含义，并避免由于属性贡献度不同而造成的聚类结果的差异。如下是聚类算法中度量样本相似性最常用的的几种距离度量方式：

曼哈顿距离，如式 2.3 所示：

$$dis(x_i, x_j) = \|x_i - x_j\| = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{id} - x_{jd}| \quad (2.3)$$

欧氏距离，如式 2.4 所示：

$$dis(x_i, x_j) = \|x_i - x_j\| = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{id} - x_{jd})^2} \quad (2.4)$$

闵可夫斯基距离，如式 2.5 所示：

$$dis(x_i, x_j) = \|x_i - x_j\| = \sqrt[q]{|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{id} - x_{jd}|^q} \quad (2.5)$$

其中，x_i和x_j表示两个样本点，q 为正整数，q=1 时即为曼哈顿距离；q=2 时即为欧氏距离。

本文结合 K-Means 算法的相似性度量方式，在欧氏距离度量方式的基础上设定加权欧氏距离，如式 2.6 所示：

$$dis(x_i, x_j) = \|x_i - x_j\| = \left(\sum_{n=1}^{d} w_n (x_{in} - x_{jn})^2\right)^{1/2}$$

$$= \sqrt{w_1 (x_{i1} - x_{j1})^2 + w_2 (x_{i2} - x_{j2})^2 + \ldots + w_r (x_{id} - x_{jd})^2} \quad (2.6)$$

其中，任意两个数据对象间的距离为$\|x_i - x_j\|$，d表示聚类变量的个数，$x_1, x_2, \ldots, x_d$为聚类变量，相对应的属性权重为$w_1, w_2, \ldots, w_d$，且$\sum_{i=1}^{d} w_i = 1$。

在实际应用中，银行客户数据的情况通常既包含数值型数据，又有少量分类型数据。针对这一情况，需要经过一定的编码才能进入模型，例如年龄等有序数据可直接进行标准差标准化然后进入模型。但是性别、归属地等无序变量无法直接进入模型，需要进行设定哑变量和 WOE（证据权重）编码的方式进行连续化处理等，例如性别变量中“男性”设定为 1，“女性”设定为 0；归属地则通过 WOE 编码计算当前分组中响应的客户和未响应客户的比值，与所有样本中这个比值的差异，并对差异值取对数进行差异排序分组，最后进行哑变量处理。

2.2.3 属性加权聚类算法流程

K-Means 算法是一种简单且实用的聚类算法，本文在其基础上进行改进，通过对研究对象赋予每个数据属性相应的权重，体现属性对聚类目标的贡献度，从而提高聚类的准确性。属性加权聚类的算法流程与传统 K-Means 算法类似，算法包含三个流程：聚类中心初始化、初始分配和再分配。首先随机生成初始聚类中心，然后根据上节中的加权欧氏距离度量方法计算每个样本到各聚类中心的距离并将其分配到距离最近的聚类中，循环迭代，直到每个聚类中的样本不再发生改变或者达到目标函数要求。

针对已确定的聚类个数k和包含d个属性变量$x_1, x_2, \ldots, x_d$的数据集S，应用基于属性加权的聚类算法详细步骤如下：

步骤 1：根据 Logistic 逐步回归系数权重设计方案，确定属性权重。数据集S中共计d个属性变量，相对应的属性权重为$w_1, w_2, \ldots, w_d$，且$\sum_{i=1}^{d} w_i = 1$，最终形成权重矩阵w，如式 2.7 所示：

$$w = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & w_d \end{bmatrix} \tag{2.7}$$

步骤 2：属性变量的归一化。由于数据量纲和取值范围的差异，为保证数据变量间的可比性，所以需要对数据进归一化，本文综合对比各种归一化方法的结果之后，最终确定最小-最大值标准化（又叫离差标准化）的结果最优，因此本文使用的是最小-最大值标准化对原始数据进行线性变换，将数值映射到[0,1]之间，转化公式如 2.8 所示，然后将标准化后的数据集与权重矩阵相乘并进行下一步操作。

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{2.8}$$

步骤 3：确定初始聚类中心。初始聚类中心的确定对聚类结果的影响很大，常见的方法是从 N 个样本数据中随机选取 k 个对象作为初始的聚类中心，可实验多次对比轮廓系数和 CH 指标取最优结果。

步骤 4：将数据对象分配到相似度最近的聚类中，并重新计算聚类中心。根据改进的加权欧氏距离度量方式分别计算每个样本到各个聚类中心的距离，将对

象分配到距离最近的聚类中，待所有对象分配完成后，重新计算 K 个聚类的中心。

步骤 5：循环比较。重复循环步骤 4 直到聚类中心不再发生变化或达到聚类目标要求，最后输出聚类结果。

2.2.4 聚类算法的评价标准

聚类是一种非常重要的无监督学习技术，对于一个数据集来说，运用不同的聚类方法可能会得到不同的聚类结果，但是哪类结果最符合我们的期望，需要一定的评价标准。一般来说，聚类算法的评价标准分为两类：一类是聚类算法的性能评价，衡量的是聚类方法本身的性能，即聚类算法对研究问题的适用性、对异常值的干扰性以及聚类的时间空间复杂度等；另一类是聚类算法的有效性评价，衡量的是聚类结果的准确性，包含聚类结果的外部质量评价、内部质量评价和相对质量评价。

（1）聚类算法的性能评价

聚类算法种类繁多，应用聚类算法时首要考虑的是这个算法适不适合，能否高效准确的完成任务，这便需对各种算法本身的性能进行研究，进而决策出最适用的聚类算法。聚类算法性能的优劣可以从很多角度衡量，大致上可以从算法的前提条件、适用条件和算法效率三方面来上看。前提条件包含是否需要提前确定聚类个数、是否需要结合相关业务经验等；适用条件包含数据类型、数据集规模、数据维度、能否处理任意形状以及异常值等；算法效率则包含算法的时间复杂度、空间复杂度和算法的准确性等。

综合上文对聚类算法的分类总结和局限性，得出如表 2.1 所示的若干具有代表性的聚类算法性能比较结果。针对本文中的银行客户数据，则可以合理性的选取最适合的方法进行聚类并优化改进，以更好的聚类性能得到更加的准确实用的结果。具体如表 2.1 所示

表 2.1 多种聚类算法的性能比较

算法名称	可伸缩性	适用数据类型	处理高维数据能力	异常数据抗干扰性	聚类形状	算法效率
K-Means	一般	数值型	较高	较低	圆形/球形	一般
K-modes	一般	分类型	较低	较高	任意形状	较高
k-prototypes	较高	混合型	较低	较低	任意形状	较高
属性加权聚类	较高	混合型	很高	较高	任意形状	较高
BIRCH	较高	数值型	较低	较低	圆形/球形	很高
CURE	较高	数值型	一般	很高	任意形状	较高
DBSCAN	一般	数值型	较低	较高	任意形状	一般
Wave-Cluster	很高	数值型	很高	较高	任意形状	很高
CLIQUE	较高	数值型	较高	较高	任意形状	较低
SOM	很高	混合型	较高	较高	任意形状	较高
GMM	很高	数值型	较低	较低	任意形状	较高

（2）聚类算法的有效性评价

聚类算法的有效性评价是对聚类结果的精确性进行衡量，其评价方式分为外部质量评价、内部质量评价和相对质量评价。针对聚类结果的外部质量评价与监督学习评价方法相似，指的是用已知结果的原有标签数据与聚类输出结果进行对比，其理想情况是具有相同类标签的数据聚合到相同的簇中，且具有不同标签的数据相分离。该方法评价指标也很多，如纯度（purity）表示的是簇内包含单个类对象的度量；如互信息（MI）和标准互信息（NMI）使用信息熵表示的是簇内包含单个类对象的另一种度量；如准确性（accuracy）衡量聚类正确的对象所占百分比；如兰德指数（RI）和调整兰德指数（ARI）衡量的是两个数据分布的吻合程度；此外，还有 F 值评价法、Rand 指数和 Jaccard 系数等等。

一般情况下，无监督聚类评估要比有监督学习评估的难度高，绝大多数情况下由于聚类结果的标签无法获得导致外部质量评价方法往往无法应用。因此，国内外很多文献中都提出了针对聚类算法的内部质量评价和相对质量评价方法对无监督的聚类结果进行评价，诸如 CH 指标、轮廓系数、戴维森堡丁指数（DBI）、邓恩指数（DVI）、Cophenetic 相关系数、Hubert's Γ 统计等等，然而这些算法可能会受诸如噪声、单调性等数据“异常”的影响，因此本文采用多指标结合的方法从紧密度、分离度、CH 指标和轮廓系数四个角度个角度共同评价银行客户聚类结果的好坏，方法原理如下：

（1）紧密度

紧密度（Compactness, CP）指的是各数据对象到簇中心的平均距离，用来衡量簇内样本点之间的是否紧凑，CP 值越低意味着类内聚类距离越近，如下式 2.9 所示：

$$\overline{CP_i} = \frac{1}{|\Omega_i|}\sum_{x_i \in \Omega_i}\|x_i - w_i\| \qquad \overline{CP} = \frac{1}{K}\sum_{k=1}^{k}\overline{CP_k} \qquad (2.9)$$

其中，$\|x_i - w_i\|$表示样本簇i内的点x_i与聚类中心w_i之间的距离，k为聚类的簇数，即聚类类别数。

（2）分离度

分离度是衡量一个簇内样本点与其他簇之间的距离是否足够的远，这里用间隔性（Separation, SP）指标来表示，指的是各聚类中心两两之间的平均距离，SP 值越高意味类间聚类距离越远，如式 2.10 所示：

$$\overline{SP} = \frac{2}{k^2-k}\sum_{i=1}^{k}\sum_{j=i+1}^{k}\left\|w_i - w_j\right\|_2 \qquad (2.10)$$

其中，w_i和w_j表示两个样本簇i和j的聚类中心，$\left\|w_i - w_j\right\|_2$表示两个样本簇聚类中心之间的距离，$k$为聚类的簇数，即聚类类别数。

（3）相对评价指标

紧密度和分离度性能比较单一化，衡量效果较为片面，紧密度没有考虑类间效果，分离度没有考虑类内效果，因此就需要能够同时衡量紧密度与分离度的整体指标来综合评价聚类的结果。本文采用相对评价指标有如下两个：

①CH 指标(Calinski-Harabasz)

CH 指标计算简单直接，通过上文得到的分离度和紧密度的比值得到，CH 指标越大代表着类自身越紧密，类与类之间越分散，其聚类结果更优。如式 2.11 所示：

$$CH = \frac{trB(k)/(k-1)}{trW(k)/(n-k)} \qquad (2.11)$$

其中，$B(k)$为类别间的协方差矩阵，$W(k)$为类别内部数据间的协方差矩阵，tr表示矩阵的迹。

②轮廓系数（Silhouette Coefficient, SC）

轮廓系数同时兼顾了聚类的凝聚度和分离度的度量，用来度量聚类结果的整体质量。该系数可以在相同原始数据的基础上评价不同算法或算法不同运行方式对聚类结果所产生的影响，所有样本的$s(i)$的均值称为聚类结果的轮廓系数，取

值范围为[-1,1]，如式 2.12 所示：

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad \text{，即} \quad s(i)=\begin{cases}1-\frac{a(i)}{b(i)}, & a(i)<b(i)\\ 0, & a(i)=b(i)\\ \frac{b(i)}{a(i)}-1, & a(i)>b(i)\end{cases} \qquad (2.12)$$

其中，$a(i)$为簇内不相似度，表示样本i到同簇其他样本的平均距离，其值越小说明样本i越应该被聚类到该簇；$b(i)$为簇间不相似度，表示样本i到其他某簇C_j的所有样本的平均距离，其值越大说明样本i越不属于其他簇。

3 银行客户细分目标与数据准备

上一章节详细介绍了基于属性加权聚类的客户细分算法理论和实现流程，本章则是对客户细分算法的实际应用奠定基础，主要包含两部分内容：第一部分是对银行客户细分的业务目标进行分析，在明确银行企业的需求和客户细分的最终目的的基础上，确定我们要进行实证分析的具体业务目标；第二部分则是对客户细分算法实现的数据进行前期处理和展示，具体可分为数据来源和指标释义、数据预处理过程和数据的基本特征统计三方面内容，其中最主要的是对于银行客户的数据预处理过程，包括数据清洗、数据转换和变量选择，最终得到模型可用的数据对象。

3.1 业务目标分析

在进行实证研究前，对业务目标有一个清晰准确的认识至关重要，它直接决定了一个项目能否实施成功以及数据挖掘结果的可用程度，然后通过数据分析与挖掘的方法解决问题、提升工作效率、用数据指导运营决策、驱动业务增长等等。客户细分的最终目的是为银行提供真正实用的建议，就是要更精准的决策出海量的银行客户中哪些是活跃的高价值客户，哪些是可放弃的低价值客户，哪些是具有成长性的潜在客户，哪些是最容易流失的 VIP 客户，哪些是易于培养的忠实客户；同时需要合理的判断出银行企业应当怎样迎合优质客户的需求，应当从什么角度扩大和吸引重点客户，应当如何根据消费倾向给客户推荐合适的产品及服务。只有解决这些问题，银行企业才能耗费更少的资源得到更好的盈利，才能实现利益最大化，为银行未来决策提供依据。

根据“二八法则”，只有核心群体才能贡献更大的价值，因此针对客户群体不同采取手段也不同。客户量庞大而银行的资源有限，当只能投入有限资源的时候，往往会选择核心群体，因此实现客户的精准化细分可以为银行业实现企业利益的最大化提供决策基础。本文对客户的细分策略分为两个阶段：客户分层与客户分群。所谓的客户分层就像我们社会阶层一样，比如按社会经济地位划分，可分为普通大众、中产、精英、富豪，而本文通过客户的 AUM 资产总额（金融总资产）这一指标把客户分为低端客户、中端客户和高端客户三层。各个层级的用户都有其不同的特点，即便低端客户中也有为银行带来高收益的客户，即便高端

客户中也有带来低收益的客户。

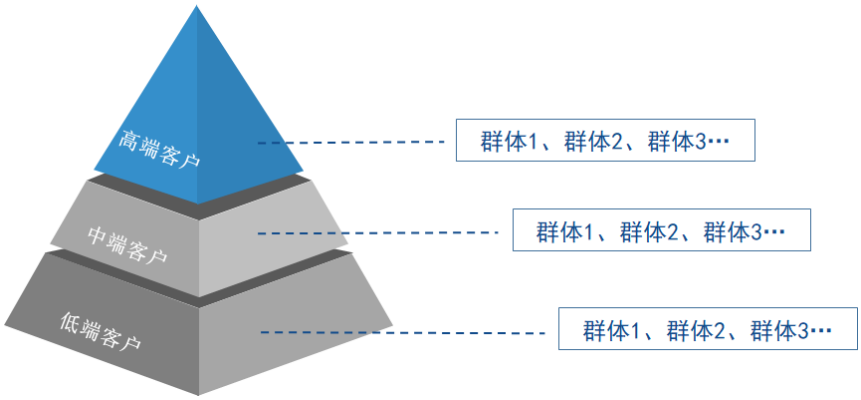


图 3.1 客户细分模型结构

如图 3.1 显示，为客户细分模型的结构，呈金字塔形，上下层之间呈递进关系，水平层之间呈依赖关系。其中，客户分层是上下结构，是基于大方向的划分，简单高效，而客户分群则是本文研究的主题，是水平结构上的一种划分。它将同一个分层内的群体通过本文改进的聚类算法继续切分，以满足更高的精细化需要。客户分层与客户分析两者相辅相成，都是客户细分模型的一部分，只有通过更加精准高效的客户细分，我们可以有目的的制定出更有针对性的运营策略，避免浪费，使客户资源高效化。

3.2 数据来源及指标释义

本文使用的是从某商业银行数据库和 CRM 系统中随机抽样的 8 万客户在该银行全年的所有交易记录及相关信息，以客户 ID 作为唯一标识，不含姓名、身份证号、手机号、家庭住址等敏感字段，由数据库和 CRM 系统的真实数据在 SQL、SAS 与 Stata 环境中，经过初步的数据采样、数据清洗、数据规约、指标构建等来构建模型（数据来源链接：http://student.saschampion.com/）。

本文主要以客户当月 AUM 资产月日均为主要研究对象，以客户为银行所带来的收益为研究目标。为了研究主题的需要，需要进行一系列的数据仓库和 CRM 系统调研及数据探索，最终确定某商业银行客户基本属性信息、客户标识信息、客户价值信息、RFM 信息、客户交易及动账最值信息五个维度，

综合筛选 628 个数据项，共计 8 万条客户数据，作为数据模型的初始目标数据集，进行银行客户分层、客户分群、客户特征分析、客户价值评估等数据挖掘任务。

由于商业银行数据指标的复杂性，这里针对常见的指标进行释义。其中最重要的指标为资产管理规模（Asset Under Management, AUM）：该指标衡量的是金融机构资产管理业务规模，是该机构当前管理客户资产的总市值。AUM 越大，说明其资产越多，客户价值更高。该指标通常以客户每个月中每天的 AUM 时点余额，取平均得到当月 AUM 月日均，作为本文主要研究对象。此外，商业银行客户行为数据，重点从客户基本属性信息、客户标识信息、客户价值信息、RFM 信息、客户交易及动账最值信息五个维度进行归纳解释。

（1）基本属性信息主要涵盖人口统计学的相关信息，如客户性别、年龄、客户持有账户数量、累计开户数目、开户年限、累计销户数目、持有本币账户以及外币账户数量等。

（2）标识信息主要包括是否个贷客户、是否关联还款、是否金普卡、是否标准白金卡、是否薪资理财、持有信用卡产品标识、持有定期存款标识、持有活期产品标识、持有基金标识、持有国债标识等。

（3）价值信息：客户持有的全部产品数量、资产总额、负债总额、客户持有的定期、活期、理财、基金、保险、国债等产品数量及交易金额等。

（4）RFM 信息：一年内客户的转入和转出等 RFM（时间、频度、金额）信息，如最近一笔取现距今天数、最近一笔转账距今天数等最近一次时间（Recency）信息；当月交易笔数、当月消费笔数、当月取现笔数和当月境外交易笔数等频率（Frequency）信息；当月交易金额、当月消费金额、当月取现金额和当月境外交易金额等金额（Monetary）信息以及出入账比例等新增信息。

（5）交易及动账最值信息：作为 RFM 信息的补充，涵盖当月、近三月和近六月内客户的最值信息，包含最大和最小消费、最大和最小取现、最大和最小动账金额和累计消费、累计取现、累计动账金额等信息。

通过客户近一年观察期数据的初步分析，对以上 5 方面因素进行细化及拆

分，共设计 5 个维度 628 个数据项作为建模基础分析宽表的要素。初始分析宽表数据项详细见附表 3.1。

3.3 数据预处理

3.3.1 数据清洗

商业银行原始客户数据中，存在着大量不完整、不一致、有噪音的数据，严重影响到对客户细分模型的执行效率，甚至可能导致挖掘结果的偏差，所以进行数据清洗是整个数据挖掘过程中不可缺少的一个环节，在实际操作中数据清洗通常会占据大半的时间，其结果质量直接关系到模型效果和最终结论。数据清洗的对象为低端客户、中端客户和高端客户三层，因此对于商业银行客户数据进行清洗时，需要预先对数据进行分层，针对每层客户的数据清洗方式也各不相同。数据清洗主要包含缺失值处理与异常值处理，也包括删除与挖掘主题无关的数据、原始数据集中重复数据和噪声数据等等。

数据清洗的方法也不尽相同，针对缺失值的处理主要依据缺失值的分布情况和缺失值所在属性的重要程度，分为三种类型：当缺失率高（>99%）时，直接删除此数据；当缺失率较高（>95%）且属性重要程度低时，直接删除此数据；其他情况含有缺失值时，利用拉格朗日插值法根据已知点建立合适的插值函数对缺失值近似替换，或直接不处理（默认为空值或 0）。针对异常值的处理，主要采用统计学的方法，利用百分位数分布和箱型图检测异常值，对于异常的点根据变量指标特征统计结果，结合业务含义，规定每一变量的极值，将超出极值范围的异常值转化为规定的极值，如 90%分位数之前都正常变化但之后数值突然变化及其异常的情况。

3.3.2 数据变换

数据清洗完成后接着要进行或同时进行数据集成、转换、规约等一系列的处理，在本文中统称为数据变换，主要包含指标加工、指标转化和衍生新指标三类。

（1）指标加工：主要包含对数值型变量的归一化和类别型变量的编码，同上文第三章节中所述，为解决由于数据量纲不同而带来的数据变量间不可比性的问题，需要对数值型数据进标准化，同时对分类型数据进行编码。

（2）指标转化：对指标数据转化成更便于统计的数据形式。转化方式包括将连续性变量转化为离散变量（将金额、笔数转换为是否等布尔值）、将非正态分布变量数据 log 转换成为均衡分布（将 AUM 值、交易金额等转换为近正态分布）、将变量转换为二分类变量以增强变量对目标变量的解释能力，将变量进行平方、开方、取对数、差分运算等等。

（3）衍生新指标：即属性变量构造，根据业务含义及数据时间窗口（数据中当前月份）衍生新的指标。

3.3.3 变量选择

模型的建立是一个反复探索、迭代、比较的过程，需进行多个环节的分析、验证，才能得到符合业务逻辑、解释度优良的模型。由于银行客户原始数据较为复杂，628 个数据项过于繁琐，很难直接适用于 logistics 回归权重设计方案，更不能直接用于聚类，因此需要预先对银行客户初始数据进行筛选指标。本文则通过基本统计分析、趋势分析、业务分析、相关性分析等方法进行变量的选择与确定，通过模型尝试、比较及业务解读确定建模方法，最终得到具有可解释性、可靠的相关变量。

（1）基本统计分析

对于变量选择的一个主要方法是对数据总体分布进行探索，包含检查数据指标缺失值、正值、零值、负值和 NULL 值的数量和比例，以及对指标特征统计及分布进行探索，查看每一变量的最小值、最大值、平均值、中位数、十等值分箱分布及十分位数分箱分布等等。比如一个变量的取值全部相同时则此变量指标无效或者一个指标变量取值与事实相反时则删除该指标。经初步探索，商业银行低端客户占据大半占比 71%，中端客户和高端客户比例近似一致，占比约 14%，三层客户分别进行最终的模型建立。

（2）趋势分析

对于连续变量，应当着重选择变量增长趋势与目标变量 AUM 值近似相同或

完全相反的变量，而变化趋势不规则或者变化趋势平缓的变量依据业务含义按需筛选；对于分类变量，选择各分类值对应的目标变量 AUM 值变化率与基准 AUM 值变化率相比波动较大的指标。

（3）业务分析

银行客户细分模型的目标重点是能够应用于实际而且符合业务要求，脱离业务进行分析如同无根之萍，没有意义。因此，我们需要根据业务指标含义和历史经验，选择取对目标变量区分能力、业务解释能力较强的变量，并消除变量间共性。最终确定的基础数据分析宽表变量详见附表 3.2。

（4）相关性分析

经过数据清洗和数据变换的初筛，以及基本统计分析、趋势分析和业务分析的第二轮筛选，相关分析则是变量选择的最重要一步。由于 Logistic 回归模型对多重共线性较为敏感，为避免两个或多个高度相关自变量同时放入模型，导致变量系数正负被扭转，因此需进行相关性分析以减少候选变量之间的相关性。针对低端客户、中端客户和高端客户三层数据分别构建若干指标的相关性矩阵，剔除那些意义高度相同的变量和与目标变量相关性极低的指标，例如 2 个指标间的相关性较强，相关系数大于阈值（根据变量总数确定，一般取 0.7），则删除其中与目标变量相关性较低的变量，保留与目标变量相关性较大的变量，且确保该指标有较强的业务解释能力。变量的选择与确定流程如图 3.2 所示：

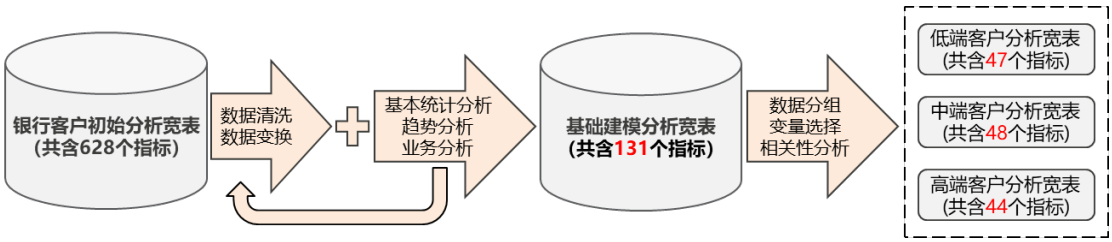


图 3.2 变量选择与确定

3.4 数据基本特征统计

银行客户细分算法模型的建立不是一蹴而就，而是通过大量的时间探索修正，其中最为耗时的便是模型指标的构建与筛选。通过对银行客户近一年的基本特征

与行为特征处理，共计得到628个数据指标，繁琐的数据项为无法使用的数据或者与模型业务目标本身不相关，因此本文对数据的处理方面采取了多个步骤环环相扣、层层筛选的方案。第一步，对含有628个指标的银行客户初始分析宽表进行数据清洗和数据变换，删除、替换与新增部分变量；第二步，对经过处理初始分析宽表数据进行基本统计分析、变量趋势分析和业务目标解读分析，得到含有131 个指标的基础建模分析宽表；第三步，对基础建模分析宽表数据进行分组，得到低端客户、中端客户和高端客户均含131个指标的三组宽表数据，并分别对三组数据重复第二步的变量选择过程；第四步，对第三步处理后的三组数据进行相关性分析，最终得到共含47个指标的低端客户分析宽表，共含48个指标的低端客户分析宽表，共含44个指标的低端客户分析宽表，最终用于客户细分数据建模。低端、中端和高端客户分析宽表数据的数据基本特征分别如附表3.3、附表3.4和附表3.5所示。

4 银行客户细分算法实现与应用

前面内容详细介绍了本文所采用的属性加权聚类算法理论和实现流程，并对需要实现的业务目标进行分析，同时详细描述了算法实现的前期数据准备过程，接下来本章则是本文的核心实践部分，重点介绍了基于属性加权聚类的客户细分算法的实现与应用。首先，对聚类算法的属性权重进行确定，通过 Logistic 逐步回归的结果实现和结果评估最终确定聚类算法所用的变量和变量权重矩阵；然后，对聚类结果的可视化展示，确定聚类的 K 值并得出传统 K-Means 聚类算法和本文改进的属性加权聚类算法可视化结果比较；其次，传统 K-Means 聚类算法和属性加权聚类算法进行有效性评价比较，证明了本文改进的聚类算法的优越性；最后，对属性加权聚类后的客户细分结果进行特征和价值分析，确定哪类客户具有明显的消费倾向、哪类客户属于重点保持客户、哪类客户属于易流失客户、哪类客户属于可撇弃的低价值客户等等。

4.1 属性权重确定

4.1.1 Logistic 逐步回归实证

由全年客户表现生成客户行为指标，以当前月份（观察点）客户 AUM 月日均为基准，观测其在未来三个月（观察期）的资产变化。若观测其在未来三个月内资产若增加 30%以上，则视为达到提升目标，定义为 1；否则视为不达标，定义为 0。针对客户在未来三个月内的达标情况作为被解释变量，建立 Logistic 预测模型，用来筛选与客户资产提升相关的重要指标，并确定权重。

具体模型根据 AUM 月日均分组如下：

第一组（低端客户）：AUM 月日均 1 万以下的客户，未来三个月内客户 AUM 月日均提升 30%以上时，定义为达标客户；

第二组（中端客户）：AUM 月日均 1 万（含）-20 万的客户，未来三个月内客户 AUM 月日均提升 30%以上时，定义为达标客户；

第三组（高端客户）：AUM 月日均 20 万（含）以上的客户，未来三个月内客户 AUM 月日均提升 30%以上时，定义为达标客户；

本次建模采样的银行客户总计 8 万人，在未来三个月内资产达标的客户约

5495 人，占比约 6.87%。探索该部分客户的差异特征是建模要解决的关键。银行客户 AUM 月日均达标情况分布如表 4.1 所示：

表 4.1 客户 AUM 月日均达标情况分布

AUM 月日均分组	客户标签	客户数(人)	客户占比(%)	达标客户数(人)	达标客户占比(%)
1w 以下	低端客户	41888	52.36	2581	6.16
1w-20w	中端客户	26614	33.27	2209	8.30
20w 以上	高端客户	11498	14.37	705	6.13
合计		80000	100.00	5495	6.87

由表 4.1 可知，本次建模客户群体的总体样本中，仅有 6.41%的客户在未来 3 个月的观察期内达标，93.59%的客户未达标，两者比例相差悬殊，若直接按原比例进入模型，将在建模中错误的将达标客户压倒性误判为“未达标”。故本文针对这种样本不平衡问题采用过采样和欠采样的分析方法，提升训练模型数据的正样本（达标客户）浓度或抽取相同比例的负样本（未达标客户），以消除模型结果过度偏向负样本的因素。同时，通过变量对不同训练集的影响分析，排除部分不稳定的变量，通过 Logistic 算法选取对目标变量有显著影响的变量，并采用逐步回归的方法按自变量对目标变量的贡献度由小到大排序，依次剔除，直至方程中没有不显著的变量可剔除为止。

经过以上各种分析方法分析后，最终分别得出进入模型的指标及其对应系数：

（1）第一组（低端客户）：

当前月份 AUM 月日均 1 万以下的客户共计 41888 人，在未来三个月实际达标（AUM 月日均提升≥30%）的人数为 2581 人，占比 6.16%。

共 10 个变量进入模型，结果如下：

$$\ln\left(\frac{p}{1-p}\right) = -6.2518 - 4.4491X_1 - 3.0252X_2 + 0.3947X_3 + 8.2399X_4 +$$

$$-13.2199X_5 - 71.3975X_6 + 0.7849X_7 - 2.7528X_8 + 1.3663X_9 + 1.3121X_{16},$$

其中 P 值为预测未来三个月 AUM 月日均提升概率。

具体模型参数如表 4.2 所示：

表 4.2 低端客户 logistic 回归估计结果-模型 1

参数	指标名称	字段类型	估计	标准误差	Pr>卡方
Intercept	常数项	—	-6.2518	0.8345	<.0001
DEP_TD_FLAG	持有定期存款标识	varchar	-4.4491	0.5124	<.0001
DEP_SA_FLAG	持有活期产品标识	varchar	-3.0252	0.833	0.0003
CUST_PRODUCT_CNT	持有的全部产品数量	integer	0.3947	0.0818	<.0001
DEP_SA_DAYAVG_BAL	本月活期存款月日均余额	decimal	8.2399	0.5483	<.0001
L3_DEP_SA_DAYAVG_BAL	近 3 月活期存款月日均余额	decimal	-13.2199	1.5878	<.0001
L3_CUST_SAVING_AVGAMT	近 3 月存款月日均金额	decimal	-71.3975	5.5996	<.0001
DEP_SA_NEW_BAL	本期本币新增余额	decimal	0.7849	0.1439	<.0001
L3_DR_CNT	近 3 月转入笔数	integer	-2.7528	0.7145	0.0001
DEP_SA_CREDIT_CNT	本期账户贷方交易次数	integer	1.3663	0.3292	<.0001
DEP_SA_MOTH_MAX_IN_AMT	当月本币单笔最大转入金额	decimal	1.3121	0.3589	0.0003

（2）第二组（中端客户）：

当前月份 AUM 月日均 1 万（含）-20 万的客户共计 26614 人，在未来三个月实际达标（AUM 月日均提升≥30%）的人数为 2209 人，占比 8.3%。

共 15 个变量进入模型，结果如下：

$$\ln\left(\frac{p}{1-p}\right) = -1.9926 + 0.5027X_1 + 0.2097X_2 - 0.6313X_3 + 0.5842X_4 - 0.6399X_5 + 0.2592X_6 + 1.1093X_7 + 0.2119X_8 + 0.9875X_9 - 0.7967X_{10} + 2.0196X_{11} + 1.3355X_{12} - 2.1329X_{13} - 0.2719X_{14} - 2.0298X_{15},$$

其中 P 值为预测未来三个月 AUM 月日均提升概率。

具体模型参数如表 4.3 所示：

表 4.3 中端客户 logistic 回归估计结果-模型 2

参数	指标名称	字段类型	估计	标准误差	Pr>卡方
Intercept	常数项	——	-1.9926	0.2658	<.0001
FUND_FLAG	持有基金标识	varchar	0.5027	0.1172	<.0001
DEP_TD_FLAG	持有定期存款标识	varchar	0.2097	0.1005	0.0369
DEP_SA_FLAG	持有活期产品标识	varchar	-0.6313	0.2466	0.0105
cr_dr_ratio	当月出入账比率	decimal	0.5842	0.1266	<.0001
L3_LG_TXN_AVG_CNT	近 3 月月均大额交易笔数	decimal	-0.6399	0.1309	<.0001
L3_DEP_CARD_CUST_CNT	近 3 月储蓄卡月均消费次数	decimal	0.2592	0.1211	0.0323
L3DEP_SA_MOTH_MAX_IN_AMT	近 3 月本币单笔最大转入金额	decimal	1.1093	0.3285	0.0007
CUST_PRODUCT_CNT	持有的全部产品数量	integer	0.2119	0.0978	0.0303
DEP_SA_BAL	本期持有本币余额	integer	0.9875	0.1211	<.0001
L6_DEP_SA_NEW_AVG_BAL	近 6 月月均本币新增余额	decimal	-0.7967	0.2299	0.0005

参数	指标名称	字段类型	估计	标准误差	Pr>卡方
CHANNEL_OTHER_IN_MAX_AMT	本期其它转入最大交易金额	decimal	2.0196	0.2958	<.0001
DEP_SA_DAYAVG_BAL	本月活期存款月日均余额	decimal	1.3355	0.198	<.0001
L6_DEP_SA_DAYAVG_BAL	近 6 月活期存款月日均余额	decimal	-2.1329	0.3828	<.0001
DEP_SA_LAST_TENURE_DAYS	活期存款最近开户距今月份	integer	-0.2719	0.0941	0.0038
L6_CUST_SAVING_AVGAMT	近 6 月存款月日均金额	decimal	-2.0298	0.325	<.0001

（3）第三组（高端客户）：

当前月份 AUM 月日均 20 万（含）以上的客户共计 11498 人，在未来三个月实际达标（AUM 月日均提升≥30%）的人数为 705 人，占比 6.13%。

共计 14 个变量进入模型，结果如下：

$$\ln\left(\frac{p}{1-p}\right) = -1.5004 + 0.38X_1 - 0.2986X_2 + 0.2182X_3 + 0.5808X_4 + 0.6027X_5 - 2.4805X_6 + 2.1227X_7 - 0.7746X_8 + 0.6786X_9 + 1.6552X_{10} + -1.3549X_{11} - 0.3452X_{12} - 0.549X_{13} + 0.3711X_{14},$$

其中 P 值为预测未来三个月 AUM 月日均提升概率。

具体模型参数如表 4.4 所示：

表 4.4 高端客户 logistic 回归估计结果-模型 3

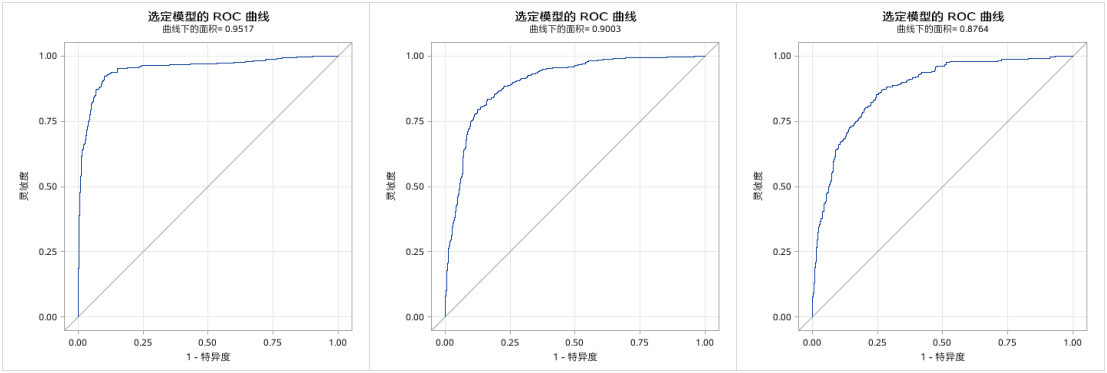
参数	指标名称	字段类型	估计	标准误差	Pr>卡方
Intercept	常数项	——	-1.5004	0.1987	<.0001
CUST_SALARY_FINANCIAL_FLAG	是否薪资理财	varchar	0.38	0.1617	0.0188
CUST_GOLD_COMMON_FLAG	是否金普卡	varchar	-0.2986	0.1006	0.003
FUND_FLAG	持有基金标识	varchar	0.2182	0.0863	0.0114
DEP_TD_FLAG	持有定期存款标识	varchar	0.5808	0.0832	<.0001
L6_CHANNEL_CTR_DTAIN_MAXAMT	近 6 月柜面异名他行转入最大交易金额	decimal	0.6027	0.082	<.0001
L6_CUST_FINA_AVGAMT	近 6 月理财月日均金额	decimal	-2.4805	0.2397	<.0001
DEP_SA_DAY_MAX_IN_AMT	单日本币单笔最大转入金额	decimal	2.1227	0.2135	<.0001
L6_CHANNEL_CTR_AVG_CNT	柜面近 6 月月均交易笔数	integer	-0.7746	0.2989	0.0095
L3_cr_dr_ratio	近 3 月出入账比率	decimal	0.6786	0.1044	<.0001
L3_DEP_SA_NEW_AVG_BAL	近 3 月月均本币新增余额	decimal	1.6552	0.2318	<.0001
DEP_SA_DAYAVG_BAL	本月活期存款月日均余额	decimal	-1.3549	0.1885	<.0001
DEP_SA_LAST_TENURE_DAYS	活期存款最近开户距今月份	integer	-0.3452	0.0782	<.0001
CUST_ACCOUNT_CNT	客户持有帐户数量	integer	-0.549	0.1191	<.0001
DEP_SA_OPEN_TENURE_DAYS	活期存款最早开户日期距今月份	integer	0.3711	0.0829	<.0001

4.1.2 Logistic 回归结果评估

为了测试我们所训练的模型是否拥有好的泛化能力，希望能够在接下来的加权聚类中取得很好的实际效果，一个最好的方法就是通过独立同分布采样的方法将原始数据划分成训练集（80%）和测试集（20%）两部分。顾名思义，训练集就是用来训练算法模型的，而测试集就是用来测试算法模型的。通过在训练集上建立模型之后，在测试集上进行模型效果评估，同时也应避免模型过拟合，以此来确定模型变量的有效性以及模型的精确性。截至目前，我们建立了三组模型：低端客户提升模型、中端客户提升模型和高端客户提升模型，其模型效果在测试集上的评估如下：

（1）ROC 曲线

ROC 曲线的全称为“接受者操作特性曲线”，对测试数据进行预测并进行 ROC 评估，目的是在尽量少的误分类基础上，尽可能多地检验出正确分类的个体。如图 4.1 所示，其横坐标为 1－特异度，表示负例覆盖率，即客户中未提升资产的客户正确地识别为未提升客户的概率；纵坐标为灵敏度，表示正例的覆盖率，即实际提升资产的客户正确地识别为提升客户的概率。其中，ROC 曲线下有条 45 度线是参照线，ROC 曲线越远离参照线则 AUC 值越大，也就显示模型的预测效果越好。其形式如图 4.1 所示：



(a)　　(b)　　(c)

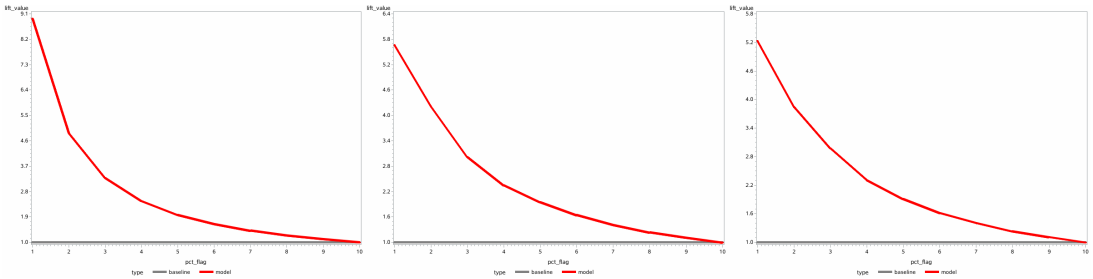
图 4.1　模型评估-ROC 曲线

如图 4.1 显示，(a)、(b)、(c)分别为低端客户、中端客户与高端客户的 ROC 曲线，三组客户模型 AUC 值分别为 0.952、0.9 和 0.876，表示任意取一个资产

达标的客户和一个资产未提升的客户，通过模型进行判定，把资产达标客户判定为正的概率为 p1，把资产未提升客户判定为正的概率为 p0，则 p1>p0 的概率分别为 0.952、0.9 和 0.876。

（2）Lift 曲线

在实际应用中，我们的首要目标并不是尽可能多地找出那些目标客户，而是提高客户的响应率。与 ROC 曲线不同的是 lift 曲线考虑分类模型的准确性，也就是使用模型获得的目标客户数量和不使用模型随机获取目标客户数量的比例，即与不用模型相比，模型的预测能力提升了多少。如图 4.2 所示 lift 曲线，其横坐标为 1-10，表示将客户按达标概率由大到小排名，并进行 10 等分；纵坐标为 lift 值，即提升度，表示使用模型各分组中实际提升的人数占比与不使用模型实际提升人数占比的倍数。



(a)　　(b)　　(c)

图 4.2　模型评估-Lift 曲线

模型中采用 Lift 曲线刻画 lift 值，图 4.4 中的(a)、(b)、(c)分别为低端客户、中端客户与高端客户的 lift 曲线。首先将客户按提升概率由大到小排名，如图 4.2 中(a)的点（1，8.93）表示使用模型时提升排名前 10%的客户中包含实际提升客户的 20.67%，是不使用模型时实际提升人数占比 2.32%效果的 8.93 倍，模型具有较高提升能力。换句话说，如果假若有 100 万的客户，平均提升率为 1%，如果对前 10%的客户进行捕获，实际上能够捕获到真实资产达标的客户 2067 人（即 100 万×1%×20.67%）。同理可得，三组模型预测概率较大的前 10%的客户达标率分别有 8.93 倍、5.66 倍和 5.23 倍提升（相比不使用模型），模型具有较高提升能力。

4.1.3 权重指标构建

经过模型的评估与验证，最终得出三组 Logistic 逐步回归结果，筛选出低端客户、中端客户和高端客户三组和 AUM 月日均提升最契合的相关变量，同时得到了与之对应的回归系数，则各个变量指标的重要性程度显而易见，由权重系数公式$w_i = |b_i| / \sum_{i=1}^{m} |b_i|$得出如表 4.5 所示的聚类变量与权重指标。

从表 4.5 可以看出，低端客户进入模型的变量总计 10 个，其中近 3 月存款月日均金额最为重要，相对其他指标来说在聚类中的贡献度应当最大；中端客户模型中最终筛选的变量为 15 个，本期其它转入最大交易金额、近 6 月活期存款月日均余额和近 6 个月存款月日均金额三个变量在模型中起决定性作用；高端客户用于聚类的特征变量 14 个，近 6 月理财月日均金额和近 3 月月均本币新增余额相对其他属性来说应当在模型中占据主导地位。

表 4.5 聚类变量与权重指标

低端客户		中端客户		高端客户	
指标名称	权重	指标名称	权重	指标名称	权重
持有定期存款标识	0.0416	持有基金标识	0.0366	是否薪资理财	0.0306
持有活期产品标识	0.0283	持有定期存款标识	0.0153	是否金普卡	0.0241
持有的全部产品数量	0.0037	持有活期产品标识	0.0460	持有基金标识	0.0176
本月活期存款月日均余额	0.0770	当月出入账比率	0.0426	持有定期存款标识	0.0468
近 3 月活期存款月日均余额	0.1236	近 3 月月均大额交易笔数	0.0466	近 6 月柜面异名他行转入最大交易金额	0.0486
近 3 月存款月日均金额	0.6676	近 3 月储蓄卡月均消费次数	0.0189	近 6 月理财月日均金额	0.1998
本期本币新增余额	0.0073	近 3 月本币单笔最大转入金额	0.0808	单日本币单笔最大转入金额	0.1710
近 3 月转入笔数	0.0257	持有的全部产品数量	0.0154	柜面近 6 月月均交易笔数	0.0624
本期账户贷方交易次数	0.0128	本期持有本币余额	0.0720	近 3 月出入账比率	0.0547
当月本币单笔最大转入金额	0.0123	近 6 月月均本币新增余额	0.0581	近 3 月月均本币新增余额	0.1334
		本期其它转入最大交易金额	0.1472	本月活期存款月日均余额	0.1092
		本月活期存款月日均余额	0.0973	活期存款最近开户距今月份	0.0278
		近 6 月活期存款月日均余额	0.1554	客户持有帐户数量	0.0442
		活期存款最近开户距今月份	0.0198	活期存款最早开户日期距今月份	0.0299
		近 6 个月存款月日均金额	0.1479		

4.2 聚类结果可视化

4.2.1 聚类类别 K 的确定

聚类类别 K 的选取方法很多，为更精准的进行聚类，本文通过对聚类个数 K 进行重复选取，从 2 类至 14 类分别对银行客户数据进行细分建模结果对比，通过 CH 指标和轮廓系趋势进行确定最佳聚类个数，在不同 K 值指标下的 CH 指标和轮廓系数分布趋势如图 4.3 所示：

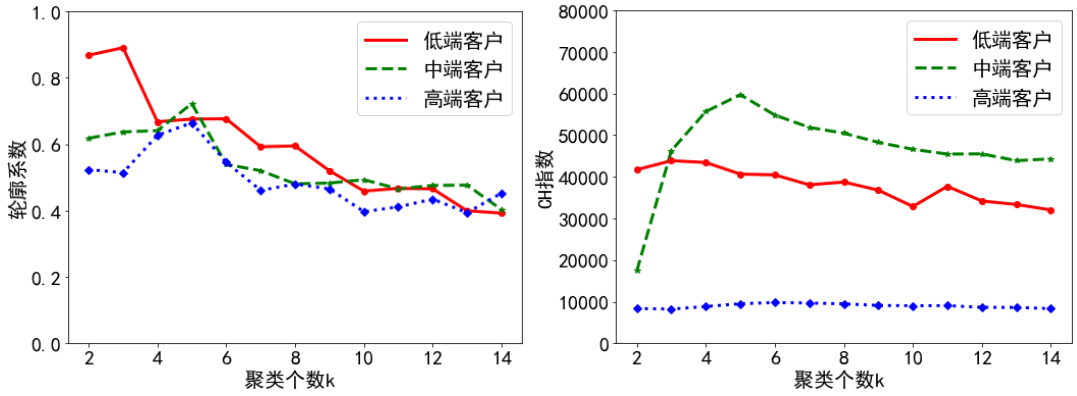


图 4.3 属性加权聚类的轮廓系数与 CH 指标趋势分布

如图 4.3 显示，左图为属性加权聚类的轮廓系数与聚类个数的趋势分布，可以很明显的看出低端客户（最上方实线）从轮廓系数的角度来看聚类效果最好，聚类个数为 3 时，轮廓系数高达 0.9，聚类效果极佳；中端客户（中间虚线）和高端客户（最下方虚线）均在 k=5 时达到至高点，轮廓系数值也较高。右图为属性加权聚类的 CH 指数与聚类个数的趋势分布，其中中端客户（最上方虚线）趋势最为明显，可确定 k=5 时聚类效果最佳；低端客户（中间实线）和高端客户（最下方虚线）分布较平缓，但可确定聚类个数的最佳范围。因此，通过综合轮廓系数和 CH 指标与多个聚类个数的趋势分析结果，最终确定低端客户最佳聚类数为 3，中端客户最佳聚类数为 5，高端客户最佳聚类数为 5。

4.2.2 K-Means 与属性加权聚类结果分布比较

依据上节得出的低端客户、中端客户和高端客户聚类变量和权重指标，分别

构建权重矩阵，并将预处理好且经过标准化后的数据与权重矩阵相乘，然后在 Python 中执行属性加权聚类算法，最终输出的属性加权聚类算法和 K-Means 聚类结果分布如表 4.4 所示：

表 4.4 K-Means 与属性加权聚类结果客户数分布比较

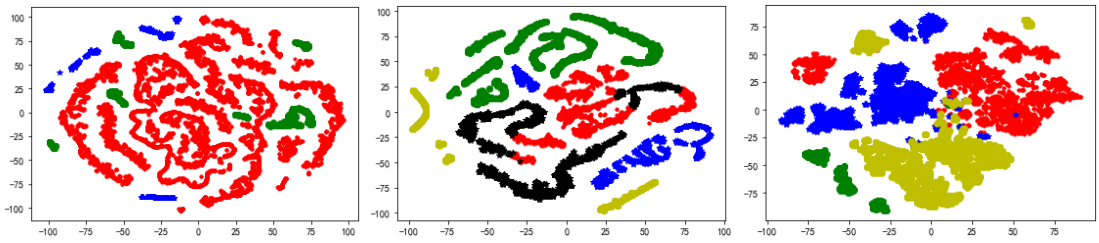
客户分组	算法分类	分群 1	分群 2	分群 3	分群 4	分群 5	总计
低端客户（AUM 月日均 1 万以下）	K-Means 聚类	29885	9688	2315			41888
	属性加权聚类	39566	1588	734	——	——	41888
	划分差异	9681	8100	1581			19362
中端客户（AUM 月日均 1 万-20 万）	K-Means 聚类	7438	7216	6587	3412	1961	26614
	属性加权聚类	7417	7237	6587	3414	1959	26614
	划分差异	21	21	0	2	2	46
高端客户（AUM 月日均 20 万以上）	K-Means 聚类	4144	3218	1816	1286	1034	11498
	属性加权聚类	4144	4072	1816	943	523	11498
	划分差异	0	854	0	343	511	1708

由表 4.4 显示，属性加权聚类和 K-Means 聚类结果多为相同划分，但聚类结果差异也较为明显，一方面显示出属性加权聚类和 K-Means 结果的有效性，另一方面显示出属性加权聚类能够根据指标的贡献度更有区分度的划分。低端客户是基数最为庞大的客户群体，同样也是划分差异最为明显的群体，两种聚类算法的划分结果相差 19362 人，占比 46.22%，从权重指标构建也可以看出，其差异原因主要受到近 3 月活期存款月日均余额和近 3 月存款月日均金额两个指标的影响，说明该指标对低端客户的资产达标和客户细分影响较为明显。中端客户属于需要重点维系的客户，经属性加权聚类后的结果与 K-Means 聚类差异较小，仅相差 46 人，占比 0.17%。高端客户作为最有价值的客户群体，为银行提供高额的收益，需要更加精准的细分，两种聚类算法的划分结果客户数相差 1708 人，差异占比 14.85%。综合而言，数量上的差异不代表类别上的差异，客户划分指标贡献度的差异可能导致不同客户划分为不同的类，从而弥补数值上的差异，探索这部分客户的差异有可能会为银行带来更高收益，因此需要更加进一步的分析。

4.2.3 K-Means 与属性加权聚类降维展示比较

依据低端客户、中端客户和高端客户 K-Means 聚类和属性加权聚类后的结

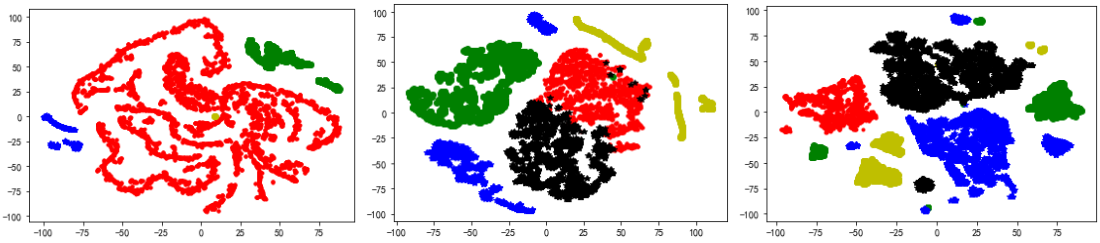
果，通过 TSNE 的方法可以用来对高维数据降维，TSNE 作为非线性降维方法，将高维映射到 2 维或 3 维的同时，原始数据分布概率尽量保证不变，从而将原始数据可视化，以便于对数据的分布有直观的了解，发现一些可能存在的规律。本文则运用 TSEN 对两种算法聚类后的结果降维并以二维的方式展示出来。因各组客户量的巨大差异，这里对三组客户均随机抽样 1 万条客户数据进行降维展示，结果如图 4.5 和图 4.6 所示，其中红色第 1 类，绿色第 2 类，蓝色表示第 3 类，黄色表示第 4 类，黑色表示第 5 类。



(a) (b) (c)

图 4.5 K-Means 聚类算法降维展示效果图

如图 4.5 显示，其中(a)、(b)、(c)分别为低端客户、中端客户与高端客户运用传统 K-Means 聚类算法得到的 TSNE 降维效果图。从二维展示图中可以明显地看出运用 K-Means 聚类可以简单对客户进行划分聚类，但是对于基数较大的低端客户数据对象分离度较差，对于中端客户的聚类内部密度效果较差，对于高端客户的类别间重叠度较高。



(a) (b) (c)

图 4.6 属性加权聚类算法降维展示效果图

如图 4.6 显示，其中(a)、(b)、(c)分别为低端客户、中端客户与高端客户运用属性加权聚类算法得到的 TSNE 降维效果图。可以明显的看到三组客户的聚类效果均非常显著，属性加权聚类算法相比于 K-Means 算法对客户数据具有更大

的区分性。尤其是图 4.6 中(b)所示的中端客户的划分聚类，虽然不可避免的存在若干点偏离聚类中心，但是属性加权聚类能够明显的区分各类，而 K-Means 算法聚出的类之间区分度不高，甚至类内间距过大。综合来看，可以初步证明属性加权聚类算法能够更加有效地对银行客户进行细分，对于具体的划分指标差异，在下文中会继续研究。

4.3 聚类算法的有效性评价

聚类作为一种无监督学习方法，评价聚类后的效果是非常有必要的，否则聚类的结果将很难被应用，在不同评价体系下评估的结果也不尽相同，本文采取多指标结合的评价标准，使用紧密度、分离度、CH 指数和轮廓系数四个指标从内部评价和相对评价角度共同衡量聚类算法的有效性。综合属性加权聚类算法和传统 K-Means 聚类算法结果计算各个评价指标，最终得出如表 4.5 所示的结果：

表 4.5 K-Means 聚类和属性加权聚类算法有效性评价对比

指标	指标名称	低端客户		中端客户		高端客户	
		K-Means	属性加权聚类	K-Means	属性加权聚类	K-Means	属性加权聚类
CP	紧密度	652.923	0.5232	461.0701	0.2077	319.082	0.2913
SP	分离度	26153512	22957.66	18969292	12407.62	2774431	2847.36
CH	CH 指数	40056.1	43877.1	41141.88	59728.88	8695.04	9775.57
SC	轮廓系数	0.8777	0.8969	0.6831	0.7220	0.6209	0.6645

紧密度、分离度、CH 指标和轮廓系数均适用于实际类别信息未知的情况，计算方式是依据标准化后的数据。正如表 4.5 显示，从紧密度指标可以很明显的看出，对属性加权聚类算法结果的紧密度明显的低于 K-Means 算法聚类结果的紧密度，说明属性加权后数据对象联系的更紧密；分离度恰好相反，属性加权聚类算法结果的分离度较之 K-Means 算法数值较小，其原因也正是由于对属性赋予权重之后，整体的数据对象距离也变小，因此，分离度与紧密度的比值，即 CH 指数的重要性不言而喻；CH 指数代表了聚类结果的分离与紧密相对程度，不管是低端客户、中端客户还是高端客户属性加权聚类 CH 指标都比 K-Means 聚类高出 10%以上；轮廓系数和 CH 指标代表的含义一致，通过不同的计算方式同时衡量样本的凝聚与分离度，其中低端客户属性加权聚类效果最佳，轮廓系数

值高达 0.9，比 K-Means 聚类高出两个百分点，中端客户和高端客户的属性加权聚类轮廓系数也较好，处于中上水平，比 K-Means 聚类高出 4 个百分点左右。

4.4 聚类结果的客户价值分析

客户细分的最终目的是为对各个细分类别的群体给与针对性的建议，为银行企业带来更高的价值利益。希望能够精准的决策出海量的银行客户中哪些是活跃的高价值客户，哪些是可放弃的低价值客户，哪些是具有成长性的潜在客户，哪些是最容易流失的 VIP 客户，哪些是易于培养的忠实客户。根据本文改进的属性加权聚类结果显示全部客群共细分为 13 个小类，在此基础上进行特征分析和客户维护发展提供建议，可以合理的判断出银行企业应当从什么角度扩大和吸引重点客户，如何根据消费倾向给客户推荐合适的产品及服务，如何提升重要发展客户的价值、稳定和延长重要潜在客户的高水平消费、防范重要挽留客户的流失并积极进行关系恢复、大力发展和维护高价值的客户并提供良好的服务等等。

4.4.1 低端客户价值分析

低端客户群体数量为 41888 人，占据总人数的 52.36%，客群基数庞大而且客户指标间的差异接近，客户发展情况参差不齐，属于最难聚类的客户群体，却也是比较有潜力的客户群体，因此找出各个聚类分群间的差异指标是主要研究目标。本文通过对属性加权聚类后的结果进行蛛网图分析，并比较各个指标在在群间的大小对某一个群体的特征进行评价分析，最后结合业务含义对各个分群之间的优势和价值进行详细刻画，并针对该群体提供针对性的服务和营销策略。针对低端客户进行属性加权聚类结果的特征如图 4.7 所示：

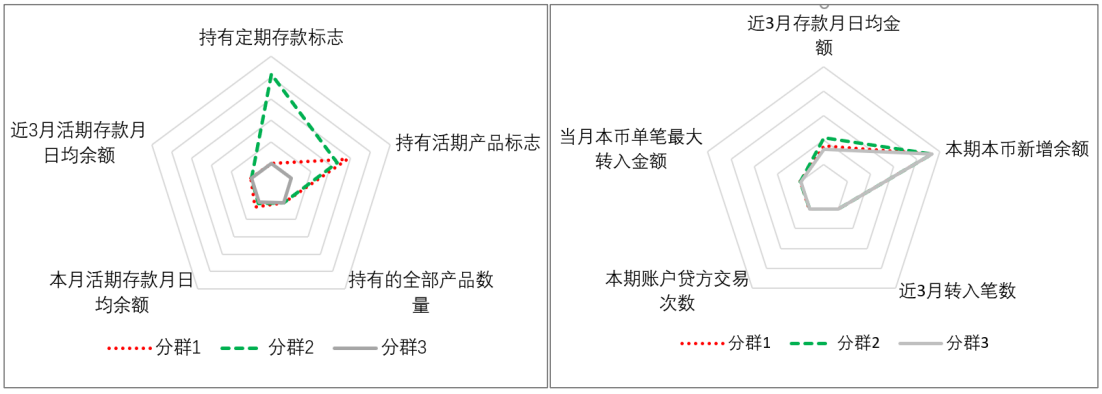


图 4.7 低端客户群特征分析图

低端客户分群 1 客户总数为 39566 人，占比 94.46%，人数较多，属于一般价值的基础客户。该类客户的优势特征为持有活期产品标识这一属性，对于该类客户应该重点关注活期产品相关业务，推荐理财、基金和存款等活期产品业务。此外，该类客户在本期本币新增余额、近 3 月存款月日均和本期活期存款月日均余额三类属性身上也有较为显著的优势，维护或者提升变动这些属性指标，可能会带来更高的收益。

低端客户分群 2 客户总数为 1588 人，占比 3.79%，该类客户属于低价值潜在客户，在持有定期存款标识这一指标上的优势最大，即说明是否拥有定期存款往往是该类客户的最显著的表现特征，对于该类客户的维护应更加关注定期存款方面。该类客户对于是否持有活期产品、本期本币新增余额和近 3 月存款月日均等方面重要性程度更大，可以适当地推荐活期产品，升级存款和本币转入服务和业务推荐。

低端客户分群 3 客户总数为 734 人，占比 1.75%，人数较少，属于低价值潜在用户。该类客户的特征优势不足，仅有本期本币新增余额和近 3 月存款月日均两个指标较为突出，在其他指标上的的重要性影响不大，因此应提升该类客户的活跃性程度而做出相应对策。

4.4.2 中端客户价值分析

中端客户群体数量为 26614 人，占据总人数的 33.27%，客群数量较大，属于需要重点发展的潜在客户群体。由于该客户群体的聚类指标较多，这里选取了

聚类划分依据最为明显的 7 类属性进行评价分析，并对每个细分群体详细的优势特征和劣势分析，最后给出针对性的客户发展建议。针对中端客户进行属性加权聚类结果的特征如图 4.8 所示：

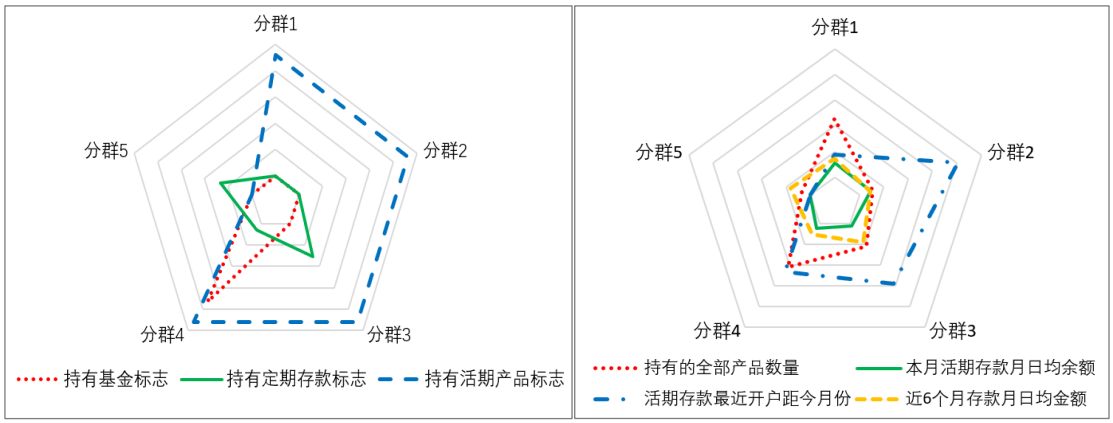


图 4.8　中端客户群特征分析图

中端客户分群 1 客户总数为 7417 人，占比 27.87%，属于重要保持客户。该类客户在持有活期产品标识和持有全部产品数量两个指标上的显著性较高，需要在这两个指标上对客户采取有利的服务与建议，以便于提升客户价值。

中端客户分群 2 客户总数为 7237 人，占比 27.19%，属于重点挽留客户客户。该类客户在持有活期产品标识和活期存款最近开户距今月份两类指标重要性较大，活期存款最近开户的时间可以代表客户的活跃状态，说明客户有一定的也无需求，因此应当重点关注该类客户的开户距今时长与活期存款状态。

中端客户分群 3 客户总数为 6587 人，占比 24.75%，该类客户属于重点保持客户。与中端客户分群 2 类似，该类客户同样在持有活期产品标识上较为显著，在活期存款最近开户距今月份上较之分群 2 较低，但是该类客户持有的全部产品数量和近 6 个月存款月日均余额两个属性重要性较高，说明该类客户处于活跃状态，经常办理相关业务。

中端客户分群 4 客户总数为 3414 人，占比 12.83%，该类客户的优势特征较多，属于需要重点维护的高价值客户。该类客户在持有基金标识和持有活期产品标识这两指标上的优势最大，即说明是否拥有基金和活期产品往往是该类客户的最显著的表现特征，对于该类客户的维护应更加关注基金和活期产品方面，为该

类客户提供更多的基金产品业务。同样的，该类客户在持有全部产品数量和活期存款最近开户距今月份两指标也较为显著。

中端客户分群 5 客户总数为 1959 人，占比 7.36%，属于低价值一般客户。该类客户在持有活期产品标识和本月活期存款月日均余额这两个属性上劣势较为明显，仅仅在持有定年期存款标识指标上略微显著，并不建议一些活期、基金和理财等产品推荐。

4.4.3 高端客户价值分析

高端客户群体数量为 11498 人，占据总人数的 14.37%，客群数量较小但是均是高价值用户，需要重点维护。本文根据客户群体指标间的差异，选取了 7 类较为显著的指标进行高价值客户的评价分析，给出针对性的客户保持和维护建议。针对高端客户进行属性加权聚类结果的特征如图 4.9 所示：

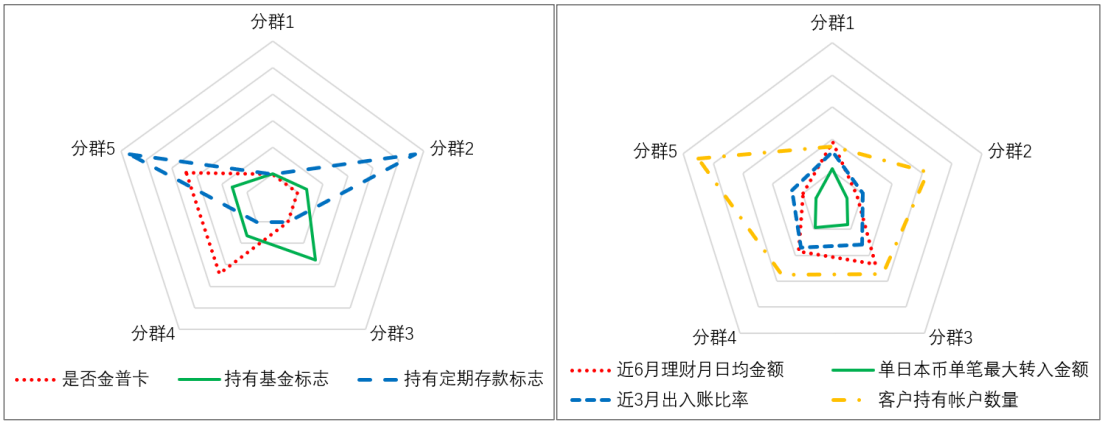


图 4.9 高端客户群特征分析图

高端客户分群 1 客户总数为 4144 人，占比 36.04%，该类客户的劣势特征较多，属于需要重点挽留的高价值客户。该类客户在是否金普卡、持有基金标识、持有定期存款标识和单日本币单笔最大转入金额等指标上较显不足，客户出入账频率低，比较不活跃，建议从推荐金普卡和提升服务和理财产品推荐等方面考虑重点挽留。

高端客户分群 2 客户总数为 4072 人，占比 35.41%，属于一般高价值客户。该类客户的价值主要在持有定期存款方面，同时客户持有账户数量也较多，建议

从定期存款转化方面实现客户价值，提升客户的活跃度。

高端客户分群 3 客户总数为 1816 人，占比 15.79%，属于需要重点维护的高价值客户，是全部客户中最具价值意义的客户。该类客户在持有基金标识、客户持有账户数量、近 6 月理财月日均金额和近 3 月出入账比率等指标上均有相当的重要性，客户活跃度较高，建议推荐金普卡、提升服务质量和提高关注度。

高端客户分群 4 客户总数为 943 人，占比 8.2%，属于重点发展客户。该类客户在是否金普卡、客户持有账户数量、近 6 月理财月日均金额和近 3 月出入账比率等指标具有较大的优势，具有较高的活跃度，客户的价值水平也较高，建议重点发展。

高端客户分群 5 客户总数为 523 人，占比 4.55%，属于需要重点保持客户。该类客户不仅在持有定期产品标识和是否金普卡指标上较为显著，而且对于客户持有账户数量和近 3 月出入账比率上更为活跃，因此需要重点保持该类客户。

5 总结与展望

本章主要对前面所有内容进行总结概括，简单扼要的总结本文所做的工作，有利于理清本文的行文思路，针对基于属性加权的聚类算法在银行客户细分中的主要步骤进行梳理和概括，并进行前景展望。

5.1 总结

本文在详细了解银行业务相关指标及概念的基础上，主要以客户 AUM 资产月日均是否提升为主要研究对象，从客户基本属性信息、客户标识信息、客户价值信息、RFM 信息、客户交易及动账最值信息五个维度，进行银行客户细分、聚类结果特征分析及客户价值分析等数据挖掘任务，以期在银行金融业务分析中给予重要的决策标准，能够运用加权聚类客户细分方法支撑产能飞跃，给予银行实际盈利收入。首先依据客户当月 AUM 月日均(金融总资产)信息，探索客户数分布，结合业务分析经验把客户分为低端客户（AUM 月日均小于 1 万）、中端客户（AUM 月日均 1 万-20 万）、高端客户（AUM 月日均大于 20 万）共计三组，主要研究内容总结如下：

（1）基于属性加权的聚类算法方案设计：本文一开始对银行客户细分技术与聚类算法进行了详细的概述，针对 K-Means 聚类的优势和不足进行全面分析，在其基础上进行改进并应用到银行客户细分中。然后针对三组客户，以客户 AUM 月日均是否提升为目标变量建立 logistic 逐步回归模型，得到与客户资产达标有关的若干变量与参数，依据回归权重设计的方法设定属性权重，改进一种加权欧氏距离的度量方式，从而得出加权聚类算法的设计方案。

（2）数据预处理及分布探索：首先，对于三组客户分别进行包含缺失值与异常值处理在内的数据清洗，同时删除与原始数据集中的无关数据、重复数据，平滑噪声数据，筛选掉与挖掘主题无关的数据等等；其次，通过指标加工、指标转化和衍生新指标等数据变换操作进行数据集成、转换、规约等一系列的处理；最后，通过基本统计分析、趋势分析、业务分析、相关性分析等方法进行变量的选择与确定，最终得到具有可解释性、可靠的相关变量。

（3）logistic 逐步回归实证结果分析、评估、验证以及权重确定：以客户 AUM 资产月日均是否提升作为目标变量，如 AUM 资产在未来三个月内提升，

定为 1，否则为 0，建立二分类的 logistic 逐步回归模型，于 SAS 中应用逐步回归的变量筛选方法，通过业务含义进行变量间的反复探索、比较与解读得到 logistic 逐步回归结果。同时，通过 ROC 曲线、AUC 指标、Lift 提升曲线和 Lift 提升值进行模型的评估与验证，得到具有可解释性、可靠的相关变量和模型参数。最终保留后的变量值与参数值依据回归权重设计的方法进行加权聚类权重矩阵设计，并得到与最终有效的显著性变量，参与到下一步应用加权聚类算法的客户细分模型中。

（4）基于属性加权聚类的银行客户细分算法实证和结果评价比较：首先，对数据预处理后的三组客户分别应用传统 K-Means 算法和改进的加权聚类算法进行客户分群，通过加权欧氏距离，将差异度最为接近的聚为一类，差异度较大的不归为一类；其次，根据 CH 指标和轮廓系数趋势分析确定最佳聚类数，最终得出客户细分后的 13 个小类，把改进的加权聚类算法与应用传统 K-Means 算法结果进行可视化比较；最后进行两类聚类算法的性能比较，和紧密度、离散度、CH 指数和轮廓系数的有效性评估对比，最终证明基于属性加权的聚类算法的优越性。

（5）客群价值分析：对于银行客户细分结果得到的三组客户各个分群进行用户基本行为特征分析，依据聚类结果、客户特征和蛛网图等分析结果，对属于同一群类别的客户进行理财、基金、定期和活期等产品倾向推荐以及潜力价值分析，以预期达到对各细分客户的潜力价值评估、个性化推荐和精准营销策略建议等。

5.2 展望

随着互联网金融时代的到来，客户数据呈爆炸式的增长，银行业纷纷向数字化科技转型，争夺客户资源，寻找优质客户，挖掘潜在客户成为银行企业竞争的潮流，但是传统的客户分类方法很难从大量数据中获取潜藏的价值和规律，因此以聚类为代表的数据挖掘算法成为客户细分的新工具。本文研究的基于属性加权的聚类算法在银行客户细分方面具有良好的应用前景，能够有效的对客户进行细分，确定每类客户的消费特征与客户价值，在很大程度上促进银行业的发展。除却本文已经完成的部分，还存在以下几点需要更加深入的研究：

（1）本文对传统 K-Means 聚类算法和改进的加权聚类算法效果进行对比分析，但是还有很多其他聚类算法如 DBSCAN、k-medoids 算法、K-prototypes 算法等，甚至改进或者发明更加精确的聚类算法，也均可加入客户细分模型结果评估的比较分析，共同衡量加权聚类算法的有效性，不过由于篇幅限制和工作量巨大，留待以后研究。

（2）本文研究的数据是银行的历史数据，具有一定的时效性与生命周期，可以在未来一段时间内内保证结果的优良性，但是随着时间的变化，模型的结果误差也会变得越来越大，因此需要每隔半年利用最新数据跑模型。此外，关于数据方面还有一些如文本数据、图像数据、音频数据等，对于数据的处理仍具有很大的挑战性。

（3）在实际应用中还存在最重要的一步——模型落地，指的是模型的上线与使用。数据模型本身没有对与错之分，只有好用与否之分。理论永远是理论，还需要对研究的模型程序封装，结合实际应用场景进行业务上的实现，例如实现一个 GUI 界面，或者设计一个前端 web 网页，又或者将模型封装继承，应用到大数据分布式系统上，实现一键导入数据、简单设定参数、一键导出结果等操作，在这一领域还有很大的研究空间。

参考文献

[1]李航. 统计学习方法[M]. 北京：清华大学出版社,2012.

[2]周志华. 机器学习[M]. 北京：清华大学出版社，2016.

[3]何晓群. 应用回归分析[M]. 北京：中国人民大学出版社,2011.

[4]林那夫, G.S.,贝里, et al. 数据挖掘技术——应用于市场营销、销售与客户关系管理[M]. 北京：清华大学出版社,2013.

[5]Smith W R. Product Differentiation and Market Segmentation as Alternative Marketing Strategies[J]. Journal of Marketing, 1956, 21(1):3-8.

[6]Lazer, William. Life style concept and marketing/toward scientific marketing[M]. Stephen Greyser, ed, Chicago: American Marketing Assn,1963:130.

[7]Haley, Russell. L, Benefit segmentation: a decision oriented research tool[J].Joumal of Marketing,1968,32(7):30-31.

[8]Hughes, A,Strategic database marketing: the masterplan for starting and managing a profitable, customer based marketing program[A].Irwin Professional, 1994.

[9]Marcus, C.,A practical yet meaningful approach to customer segmentation[J].The Journal of Consumer Marketing,1998,15(5):494.

[10]Albrecht Söllner,Mario Rese. Market segmentation and the structure of competition: applicability of the strategic group concept for an improved market segmentation on industrial markets[J]. Journal of Business Research,2001,51(1).

[11]陈明亮. 客户价值细分与保持策略研究[J].成组技术与生产现代化, 2001,(4): 23-27.

[12]张国方,金国栋. 客户细分理论及应用策略研究[J].华中科技大学学报(社会科学版),2003(03):101-104.

[13]王扶东,马玉芳. 基于数据挖掘的客户细分方法的研究[J].计算机工程与应用,2011,47(04):215-218.

[14]徐翔斌,王佳强,涂欢,穆明. 基于改进 RFM 模型的电子商务客户细分[J].计算机应用,2012,32(05):1439-1442.

[15]廉亦璇. 城市商业银行客户细分理论实践及其改进研究[D].中央民族大学,2013.

[16]许荻迪. 客户细分理论文献综述:细分维度及其在银行业的应用[J].经济师,2015(09):176-179.

[17]刘英姿,吴昊. 客户细分方法研究综述[J].管理工程学报,2006(01):53-57.

[18]Chou P B , Grossman E , Gunopulos D , et al. Identifying prospective customers[C]// Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000.

[19]Wells, William, Tigert, Doug, Actitives, interests and opinion[J].Joumal of Advertisting Research,August1971,11:27-35.

[20]林盛,肖旭. 基于 RFM 的电信客户市场细分方法[J].哈尔滨工业大学学报,2006(05):758-760.

[21]蔡玖琳,张磊,张秋三. 一种基于数据挖掘的零售业客户细分方法研究[J].重庆工商大学学报(自然科学版),2015,32(02):43-48.

[22]施荣晗,郑良琳. 商业银行金融产品的客户行为细分——基于莆田光大银行 ETC 客户数据[J].重庆城市管理职业学院学报,2018,18(01):36-39+43.

[23]郑琦. 利益细分变量研究与消费者市场细分[J].南开管理评论,2000(04):60-63.

[24]綦欣德. 客户细分方法新视角[J].商业时代,2013(26):31-33.

[25]曾小青,徐秦,张丹,林大瀚. 基于消费数据挖掘的多指标客户细分新方法[J].计算机应用研究,2013,30(10):2944-2947.

[26]王颖晖. 知识服务业背景下客户市场细分的聚类方法研究[J].统计与决策,2009(18):24-26.

[27]向昆竹,黄凯,侯皓文. 基于大数据的客户细分方法研究[J].科技风,2019(20):230-231.

[28]李卫军. K-Means 聚类算法的研究综述[J]. 现代计算机(专业版),2014(8):85-89.

[29]Huang Z . Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values[J]. Data Mining and Knowledge Discovery, 1998, 2(3):283-304.

[30]Huang Z, Ng M K. A fuzzy k-modes algorithm for clustering categorical data[J]. IEEE Transactions on Fuzzy Systems, 1999, 7(4):446-452.

[31]Zhi Z, Gong M, Ma J, et al. Unsupervised evolutionary clustering algorithm for

mixed type data[C]. Evolutionary Computation. 2010.

[32]Hsu C C , Chen Y C . Mining of mixed data with application to catalog marketing[J]. Expert Systems with Applications, 2007, 32(1):12-23.

[33]Hsu C C, Huang Y P. Incremental clustering of mixed data based on distance hierarchy[J]. Expert Systems with Applications, 2008, 177(20):4474-4492.

[34]Kim Y , Shim K , Kim M S , et al. DBCURE-MR: An efficient density-based clustering algorithm for large data using MapReduce[J]. Information Systems, 2014, 42:15-35.

[35]张晓峰. 一种基于属性加权的不确定 K-means 聚类算法[C].第 26 届中国数据库学术会议论文集,2009:515-519.

[36]陈韡,王雷,蒋子云. 基于 K-prototypes 的混合属性数据聚类算法[J].计算机应用,2010,30(08):2003-2005+2110.

[37]熊平,顾霄. 基于属性权重最优化的 K-Means 聚类算法[J]. 微电子学与计算机,2014(04):46-49.

[38]赵兴旺,梁吉业. 一种基于信息熵的混合数据属性加权聚类算法[J]. 计算机研究与发展,2016,53(05):1018-1028.

[39]黄晓辉,王成,熊李艳等. 一种集成簇内和簇间距离的加权 K-Means 聚类方法[J/OL]. 计算机学报,2019:1-15.

[40]Zakrzewska D, Murlewski J. Clustering Algorithms for Bank Customer Segmentation[C].International Conference on Intelligent Systems Design & Applications. IEEE Computer Society, 2005.

[41]孙晓霞. 聚类分析在客户细分领域的应用研究[D].西北大学,2006.

[42]花海洋,赵怀慈. 聚类算法在银行客户细分中的应用[J].计算机工程,2008,34(24):37-39.

[43]樊宁. K 均值聚类算法在银行客户细分中的研究[J].计算机仿真,2011,28(03):369-372.

[44]瞿小宁. K 均值聚类算法在商业银行客户分类中的应用[J].计算机仿真,2011,28(06):357-360.

[45]秦秀洁. 数据挖掘在银行客户关系管理中的应用研究[D].华南理工大学,2014.

[46]樊仙仙. 基于聚类分析的 H 银行客户分群及营销策略研究[D].华东理工大学,2016.

[47]李涛. 商业银行客户细分及实践研究[C]. 2016 年第二届今日财富论坛论文集,2016:23-24.

[48]于化龙,韩雪峰. 基于改进 K 均值聚类的银行客户分类算法[J]. 湘潭大学自然科学学报, 2018, 146(03):129-132.

[49]Boone D S , Roehm M . Retail segmentation using artificial neural networks[J]. International Journal of Research in Marketing, 2002, 19(3):287-301.

[50]Shin H W , Sohn S Y . Segmentation of stock trading customers according to potential value[J]. Expert Systems with Applications, 2004, 27(1):27-33.

[51]Golsefid S M , Ghazanfari M , Alizadeh S . Customer Segmentation in Foreign Trade based on Clustering Algorithms Case Study: Trade Promotion Organization of Iran[J]. International Journal of Computer, Information & Systems Science, 2007(3).

[52]Jeff Zethmayr,Ramandeep Singh Makhija. Six unique load shapes: A segmentation analysis of Illinois residential electricity consumers[J]. The Electricity Journal,2019,32(9).

[53]Hansi Chen,Lei Zhang,Xuening Chu,Bo Yan. Smartphone customer segmentation based on the usage pattern[J]. Advanced Engineering Informatics,2019,42.

[54]李鑫鑫. 聚类算法在电子商务客户细分中的应用研究[D].中国海洋大学,2012.

[55]陈治平,胡宇舟,顾学道. 聚类算法在电信客户细分中的应用研究[J].计算机应用,2007(10):2566-2569+2577.

[56]武森,程锴,陈凤洁. 聚类分析在电信客户细分中的应用[J].技术经济与管理研究,2008(01):10-12.

[57]张利利,马艳琴. 基于数据挖掘技术的航空客户流失与细分研究及 R 语言程序实现[J].数学的实践与认识,2019,49(06):134-142.

[58]汪永旗,王惠娇. 旅游大数据的 MapReduce 客户细分应用[J].华侨大学学报(自然科学版),2015,36(03):292-296.

[59]张良均,王路,谭立云等. Python 数据分析与挖掘实战[M]. 北京：机械工业出版社,2016,32-113.

附　录

附表 3.1　初始数据分析宽表字段部分截图

序号	一级指标	二级指标	英文名称	字段类型
1	客户基本属性信息	客户号	CUST_ID	long
2		开户机构	OPEN_ORG_NUM	int
3		证件类型	IDF_TYP_CD	str4
4		性别	GENDER	str1
5		年龄	AGE	int
6		是否达标	target_flag	byte
7		个贷标识	LOAN_FLAG	str1
8		持有外币账户数量	DEP_SA_FGCR_ACCOUNT_CI	int
9		持有本币账户数量	DEP_SA_ACCOUNT_CNT	int
10		信用卡最近开户时长	LAST_OPEN_TENURE_DAYS	byte
11		客户持有帐户数量	CUST_ACCOUNT_CNT	int
12		累计开户数目	CUST_ACCOUNT_OPEN_CNT	int
13		累计销户数目	CUST_ACCOUNT_LOST_CNT	int
14	客户标识信息	持有定期存款标识	DEP_TD_FLAG	str1
15		持有国债标识	BOND_FLAG	str1
16		持有活期产品标识	DEP_SA_FLAG	str1
17		持有货币型基金标识	C_FUND_FLAG	str1
18		持有基金标识	FUND_FLAG	str1
19		持有记账式国债标识	BK_BOND_FLAG	str1
20		持有偏股型基金标识	S_FUND_FLAG	str1
21		持有偏债型基金标识	D_FUND_FLAG	str1
22		持有凭证式国债标识	CER_BOND_FLAG	str1
23		持有信用卡产品标识	CRED_FLAG	str1
24		是否白金理财卡	CUST_PLATINUM_FINANCI	str1
25		是否标准白金卡	CUST_STAD_PLATINUM_FL	str1
26		是否关联还款	RELATED_REPAY_FLAG	str1
27		是否国际金卡	CUST_INTERNATIONAL_GO	str1
28		是否国际普卡	CUST_INTERNATIONAL_CO	str1
29		是否国际银卡	CUST_INTERNATIONAL_SI	str1
30		是否国际钻石卡	CUST_INTERNATIONAL_DI	str1
31		是否豪华白金卡	CUST_LUXURY_PLATINUM_I	str1
32		是否金普卡	CUST_GOLD_COMMON_FLAG	str1
33		是否美元卡	CUST_DOLLER_FLAG	str1
34		是否全额还款	TOT_REPAY_FLAG	str1
35		是否商务卡	CUST_BUSINESS_FLAG	str1

序号	一级指标	二级指标	英文名称	字段类型
43	客户价值信息	客户AUM（ECIF8种金融资产	CUST_AUM	double
44		客户持有的全部产品数量（	CUST_PRODUCT_CNT	byte
45		国债数量	guozhai_cnt	byte
46		国债金额	CUST_NADEBT_AMT	double
47		基金金额	CUST_FOND_AMT	double
48		理财金额	CUST_FINA_AMT	double
49		资产总额	CUST_ASSET_AMT	double
50		银保通金额	CUST_YBT_AMT	double
51		私人银行理财金额	CUST_PRIFINA_AMT	double
52		贵金属金额	CUST_METAL_AMT	double
53		负债总额	CUST_DEBT_AMT	double
54		客户保有期限	CUST_TENURE_MONTHS	byte
55		当月新增存款账户数	DEP_SA_NEW_ACCOUNT_CNT	byte
56		当月存款账户总数	DEP_SA_SUMACCOUNT_CNT	int
57		贷款账户月余额	G_OS_PRCP_SUM	double
58		存款金额	CUST_SAVING_AMT	double
59		本月活期存款月日均余额	DEP_SA_DAYAVG_BAL	double
60		近3个月月均外币新增余额	L3_DEP_SA_FGCR_NEW_AVG	double
61		近3个月月均持有外币余额	L3_DEP_SA_FGCR_BAL	double
62		近3个月月均持有本币余额	L3_DEP_SA_BAL	double
63		近3个月月均本币新增余额	L3_DEP_SA_NEW_AVG_BAL	double
64		近3个月月均新增存款账户数	L3_DEP_SA_NEW_ACCOUNT	byte
65		近3个月月均存款账户总数	L3_DEP_SA_AVG_ACCOUNT	int
66		近3个月银保通月日均金额	L3_CUST_YBT_AVGAMT	double
67		近3个月私人银行理财月日均	L3_CUST_PRIFINA_AVGAMT	double
68		近3个月私人银行撮合委托贷	L3_CUST_PRIDEPT_AVGAMT	long
69		近3个月内贷款账户月均余额	OS_PRCP_SUM_THREE	double
70		近3个月理财月日均金额	L3_CUST_FINA_AVGAMT	double
71		近3个月客户月均资产总计	L3_CUST_ASSET_AVG_AMT	double
72		近3个月客户月均负债总计	L3_CUST_DEBT_AVG_AMT	double
73		近3个月客户AUM平均值	L3_CUST_AVG_AUM	double
74		近3个月基金月日均金额	L3_CUST_FOND_AVGAMT	double
75		近3个月活期存款月日均余额	L3_DEP_SA_DAYAVG_BAL	double
76		近3个月国债月日均金额	L3_CUST_NADEBT_AVGAMT	double
77		近3个月贵金属月日均金额	L3_CUST_METAL_AVGAMT	double

附表 3.2　基础数据分析宽表字段汇总

序号	一级指标	二级指标	英文名称	字段类型
1	客户基本属性信息	客户号	CUST_ID	varchar
2		客户分组	cus_group	varchar
3		是否达标	target_flag	varchar
4		开户机构	OPEN_ORG_NUM	integer
5		性别	gender	varchar
6		个贷标识	LOAN_FLAG	varchar
7		持有外币账户数量	DEP_SA_FGCR_ACCOUNT_CNT	integer
8		持有本币账户数量	DEP_SA_ACCOUNT_CNT	integer
9		信用卡最近开户时长	LAST_OPEN_TENURE_DAYS	integer
10		客户持有帐户数量	CUST_ACCOUNT_CNT	integer
11		累计开户数目	CUST_ACCOUNT_OPEN_CNT	integer
12		累计销户数目	CUST_ACCOUNT_LOST_CNT	integer
13	客户标识信息	持有定期存款标识	DEP_TD_FLAG	varchar
14		持有国债标识	BOND_FLAG	varchar
15		持有活期产品标识	DEP_SA_FLAG	varchar

序号	一级指标	二级指标	英文名称	字段类型
16		持有货币型基金标识	C_FUND_FLAG	varchar
17		持有基金标识	FUND_FLAG	varchar
18		持有信用卡产品标识	CRED_FLAG	varchar
19		是否标准白金卡	CUST_STAD_PLATINUM_FLAG	varchar
20		是否关联还款	RELATED_REPAY_FLAG	varchar
21		是否金普卡	CUST_GOLD_COMMON_FLAG	varchar
22		是否美元卡	CUST_DOLLER_FLAG	varchar
23		是否薪资理财	CUST_SALARY_FINANCIAL_FLAG	varchar
24	客户价值信息	客户当月 AUM 月日均	CUST_AUM	decimal
25		持有的全部产品数量	CUST_PRODUCT_CNT	integer
26		基金金额	CUST_FOND_AMT	decimal
27		理财金额	CUST_FINA_AMT	decimal
28		资产总额	CUST_ASSET_AMT	decimal
29		当月存款账户总数	DEP_SA_SUMACCOUNT_CNT	integer
30		存款金额	CUST_SAVING_AMT	decimal
31		本月活期存款月日均余额	DEP_SA_DAYAVG_BAL	decimal
32		近 3 个月月均持有本币余额	L3_DEP_SA_BAL	decimal
33		近 3 个月月均本币新增余额	L3_DEP_SA_NEW_AVG_BAL	decimal
34		近 3 个月月均存款账户总数	L3_DEP_SA_AVG_ACCOUNT_CNT	integer
35		近 3 个月理财月日均金额	L3_CUST_FINA_AVGAMT	decimal
36		近 3 个月客户月均资产总计	L3_CUST_ASSET_AVG_AMT	decimal
37		近 3 个月客户 AUM 平均值	L3_CUST_AVG_AUM	decimal
38		近 3 个月活期存款月日均余额	L3_DEP_SA_DAYAVG_BAL	decimal
39		近 3 个月存款月日均金额	L3_CUST_SAVING_AVGAMT	decimal
40		近 6 个月月均持有本币余额	L6_DEP_SA_BAL	decimal
41		近 6 个月月均本币新增余额	L6_DEP_SA_NEW_AVG_BAL	decimal
42		近 6 个月月均存款账户总数	L6_DEP_SA_AVG_ACCOUNT_CNT	integer
43		近 6 个月理财月日均金额	L6_CUST_FINA_AVGAMT	decimal
44		近 6 个月客户月均资产总计	L6_CUST_ASSET_AVG_AMT	decimal
45		近 6 个月客户 AUM 平均值	L6_CUST_AVG_AUM	decimal
46		近 6 个月活期存款月日均余额	L6_DEP_SA_DAYAVG_BAL	decimal
47		近 6 个月存款月日均金额	L6_CUST_SAVING_AVGAMT	decimal
48		本期持有本币余额	DEP_SA_BAL	decimal
49		本期本币新增余额	DEP_SA_NEW_BAL	decimal
50	RFM 信息	活期存款最早开户日期距今月份	DEP_SA_OPEN_TENURE_DAYS	integer
51		活期存款最近开户距今月份	DEP_SA_LAST_TENURE_DAYS	integer
52		活期存款平均开户时长	DEP_SA_AVG_TENURE_DAYS	integer
53		近 3 个月转入金额	L3_DR_AMT	decimal
54		近 3 个月转入笔数	L3_DR_CNT	integer
55		近 3 个月转出金额	L3_CR_AMT	decimal
56		近 3 个月转出笔数	L3_CR_CNT	integer
57		近 3 个月账户借方月均交易金额	L3DEP_SA_DEBIT_AMT	decimal

序号	一级指标	二级指标	英文名称	字段类型
58		近 3 个月账户借方月均交易次数	L3DEP_SA_DEBIT_CNT	integer
59		近 3 个月账户贷方月均交易金额	L3DEP_SA_CREDIT_AMT	decimal
60		近 3 个月账户贷方月均交易次数	L3DEP_SA_CREDIT_CNT	integer
61		近 3 个月跨行非同名转入月均金额	L3_CUST_DIFF_IN_AMT	decimal
62		近 3 个月跨行非同名转出月均金额	L3_CUST_DIFF_OUT_AMT	decimal
63		近 3 个月储蓄卡月均消费金额	L3_DEP_CARD_CUST_AMT	decimal
64		近 3 个月储蓄卡月均消费次数	L3_DEP_CARD_CUST_CNT	integer
65		近 6 个月账户借方月均交易金额	L6DEP_SA_DEBIT_AMT	decimal
66		近 6 个月账户借方月均交易次数	L6DEP_SA_DEBIT_CNT	integer
67		近 6 个月账户贷方月均交易金额	L6DEP_SA_CREDIT_AMT	decimal
68		近 6 个月账户贷方月均交易次数	L6DEP_SA_CREDIT_CNT	integer
69		近 6 个月跨行非同名转入月均金额	L6_CUST_DIFF_IN_AMT	decimal
70		近 6 个月跨行非同名转出月均金额	L6_CUST_DIFF_OUT_AMT	decimal
71		近 6 个月储蓄卡月均消费金额	L6_DEP_CARD_CUST_AMT	decimal
72		近 6 个月储蓄卡月均消费次数	L6_DEP_CARD_CUST_CNT	integer
73		自助设备近 6 个月月均交易金额	L6_CHANNEL_AUTO_AVG_AMT	decimal
74		自助设备近 6 个月月均交易笔数	L6_CHANNEL_AUTO_AVG_CNT	decimal
75		网络银行近 3 个月月均交易金额	L3_CHANNEL_INTER_AVG_AMT	decimal
76		网络银行近 3 个月月均交易笔数	L3_CHANNEL_INTER_AVG_CNT	decimal
77		网络银行近 6 个月月均交易金额	L6_CHANNEL_INTER_AVG_AMT	decimal
78		网络银行近 6 个月月均交易笔数	L6_CHANNEL_INTER_AVG_CNT	decimal
79		近 3 个月月均大额交易金额	L3_LG_TXN_AVG_AMT	decimal
80		近 3 个月月均大额交易笔数	L3_LG_TXN_AVG_CNT	decimal
81		其它近 3 个月月均交易金额	L3_CHANNEL_OTHER_AVG_AMT	decimal
82		其它近 3 个月月均交易笔数	L3_CHANNEL_OTHER_AVG_CNT	decimal
83		本期大额交易笔数	LG_TXN_CNT	integer
84		本期大额交易金额	LG_TXN_AMT	decimal
85		本期其它贷方交易笔数	CHANNEL_OTHER_CREDIT_CNT	integer
86		本期其它贷方交易金额	CHANNEL_OTHER_CREDIT_AMT	decimal
87		本期其它借方交易笔数	CHANNEL_OTHER_DEBIT_CNT	integer
88		本期其它借方交易金额	CHANNEL_OTHER_DEBIT_AMT	decimal
89		本期账户贷方交易次数	DEP_SA_CREDIT_CNT	integer
90		本期账户贷方交易金额	DEP_SA_CREDIT_AMT	decimal
91		本期账户借方交易次数	DEP_SA_DEBIT_CNT	integer
92		本期账户借方交易金额	DEP_SA_DEBIT_AMT	decimal
93		本月储蓄卡消费次数	DEP_CARD_CUST_CNT	integer
94		本月储蓄卡消费金额	DEP_CARD_CUST_AMT	decimal
95		本月转出笔数	CR_CNT	integer
96		本月转出金额	CR_AMT	decimal
97		本月转入笔数	DR_CNT	integer
98		本月转入金额	DR_AMT	decimal
99		当月本币转账存款金额	DEP_SA_TRSP_DEP_AMT	decimal

序号	一级指标	二级指标	英文名称	字段类型
100		当月本币转账取款金额	DEP_SA_TRSP_WITD_AMT	decimal
101		柜面近 3 个月月均交易笔数	L3_CHANNEL_CTR_AVG_CNT	decimal
102		柜面近 3 个月月均交易金额	L3_CHANNEL_CTR_AVG_AMT	decimal
103		柜面近 6 个月月均交易笔数	L6_CHANNEL_CTR_AVG_CNT	decimal
104		柜面近 6 个月月均交易金额	L6_CHANNEL_CTR_AVG_AMT	decimal
105		近 6 个月本币单笔最大转出金额	L6DEP_SA_MOTH_MAX_OUT_AMT	decimal
106		近 6 个月本币单笔最大转入金额	L6DEP_SA_MOTH_MAX_IN_AMT	decimal
107		近 6 月柜面异名他行转入月均交易笔数	L6_CHANNEL_CTR_DTAIN_AVGCNT	decimal
108		近 6 月柜面异名他行转入月均交易金额	L6_CHANNEL_CTR_DTAIN_AVGAMT	decimal
109		近 6 个月月均大额交易笔数	L6_LG_TXN_AVG_CNT	decimal
110		近 6 个月月均大额交易金额	L6_LG_TXN_AVG_AMT	decimal
111		其它近 6 个月月均交易笔数	L6_CHANNEL_OTHER_AVG_CNT	decimal
112		其它近 6 个月月均交易金额	L6_CHANNEL_OTHER_AVG_AMT	decimal
113	客户交易及动账最值信息	当月出入账比率	cr_dr_ratio	decimal
114		近 3 月出入账比率	L3_cr_dr_ratio	decimal
115		本期其它转入最大交易金额	CHANNEL_OTHER_IN_MAX_AMT	decimal
116		本期其它转入最小交易金额	CHANNEL_OTHER_IN_MIN_AMT	decimal
117		单日本币单笔最大转入金额	DEP_SA_DAY_MAX_IN_AMT	decimal
118		单日本币单笔最大转出金额	DEP_SA_DAY_MAX_OUT_AMT	decimal
119		当月本币单笔最大转入金额	DEP_SA_MOTH_MAX_IN_AMT	decimal
120		当月本币单笔最大转出金额	DEP_SA_MOTH_MAX_OUT_AMT	decimal
121		其它转入 3 个月内最大交易金额	L3_CHANNEL_OTHER_IN_MAX_AMT	decimal
122		其它转出 3 个月内最大交易金额	L3_CHANNEL_OTHER_OUT_MAX_AMT	decimal
123		近 3 月单日本币单笔最大转入金额	L3DEP_SA_DAY_MAX_IN_AMT	decimal
124		近 3 月单日本币单笔最大转出金额	L3DEP_SA_DAY_MAX_OUT_AMT	decimal
125		近 3 月本币单笔最大转入金额	L3DEP_SA_MOTH_MAX_IN_AMT	decimal
126		近 3 月本币单笔最大转出金额	L3DEP_SA_MOTH_MAX_OUT_AMT	decimal
127		近 6 月柜面异名他行转入最大交易金额	L6_CHANNEL_CTR_DTAIN_MAXAMT	decimal
128		其它转入近 6 月最大交易金额	L6_CHANNEL_OTHER_IN_MAX_AMT	decimal
129		其它转出近 6 月最大交易金额	L6_CHANNEL_OTHER_OUT_MAX_AMT	decimal
130		近 6 月单日本币单笔最大转入金额	L6DEP_SA_DAY_MAX_IN_AMT	decimal
131		近 6 月单日本币单笔最大转出金额	L6DEP_SA_DAY_MAX_OUT_AMT	decimal

附表 3.3 低端客户建模指标基本统计描述

ID	Variable	Type	Min	Max	Mean	Std.Dev.
1	gender	varchar	0	2	1.4196	0.5294
2	LOAN_FLAG	varchar	0	1	0.1357	0.3425
3	DEP_TD_FLAG	varchar	0	1	0.0379	0.1910
4	DEP_SA_FLAG	varchar	0	1	0.9758	0.1538
5	CRED_FLAG	varchar	0	1	0.1457	0.3528
6	CUST_STAD_PLATINUM_FLAG	varchar	0	1	0.0407	0.1976

ID	Variable	Type	Min	Max	Mean	Std.Dev.
7	RELATED_REPAY_FLAG	varchar	0	1	0.0284	0.1661
8	CUST_GOLD_COMMON_FLAG	varchar	0	1	0.1114	0.3147
9	CUST_DOLLER_FLAG	varchar	0	1	0.0926	0.2899
10	CUST_SALARY_FINANCIAL_FLAG	varchar	0	1	0.0686	0.2527
11	FUND_FLAG	varchar	0	1	0.0382	0.1916
12	C_FUND_FLAG	varchar	0	2	0.0390	0.1977
13	LAST_OPEN_TENURE_DAYS	integer	0	85	2.4050	7.3328
14	L3_DR_CNT	integer	0	1015	0.2730	6.2188
15	CUST_PRODUCT_CNT	integer	0	13	1.4431	1.9261
16	CUST_ACCOUNT_LOST_CNT	integer	0	260	1.7296	7.1831
17	DEP_SA_LAST_TENURE_DAYS	integer	0	6257	1905.696	1129.719
18	L6_DEP_SA_AVG_ACCOUNT_CNT	integer	0	235	1.6672	1.9636
19	DEP_SA_FGCR_ACCOUNT_CNT	integer	0	29	0.3046	0.8812
20	DEP_SA_DEBIT_CNT	integer	0	54560	3.3793	266.8027
21	DEP_SA_CREDIT_CNT	integer	0	1455	2.2881	9.6961
22	CHANNEL_OTHER_CREDIT_CNT	integer	0	128	1.2991	1.4745
23	L3_CUST_AVG_AUM	decimal	0	3136444	4258.202	32139.77
24	L6_CUST_AVG_AUM	decimal	251.7283	6387140	11408.61	109894.70
25	L3_CUST_SAVING_AVGAMT	decimal	0	2728030	3445.05	19594.58
26	L3_CUST_ASSET_AVG_AMT	decimal	0	3136444	4258.202	32139.77
27	L6_CUST_ASSET_AVG_AMT	decimal	251.7283	6387140	11408.61	109894.70
28	L3_DEP_SA_BAL	decimal	0	3471871	2782.203	18464.82
29	L6_DEP_SA_BAL	decimal	0	1947334	3854.626	22482.03
30	DEP_SA_NEW_BAL	decimal	-199896	200325.4	89.0207	4825.5260
31	L3_DEP_SA_NEW_AVG_BAL	decimal	-3333526	66799.3	-787.4459	25541.83
32	L6_DEP_SA_NEW_AVG_BAL	decimal	-800979	32943.66	-549.5086	8953.789
33	DEP_SA_MOTH_MAX_IN_AMT	decimal	0	5000000	5786.464	71754.92
34	L3DEP_SA_DAY_MAX_IN_AMT	decimal	0	23900000	21594.21	248146.20
35	L3DEP_SA_MOTH_MAX_IN_AMT	decimal	0	23900000	18137.18	199329.50
36	L3DEP_SA_DAY_MAX_OUT_AMT	decimal	0	25000000	23803.87	278689.70
37	DEP_SA_CREDIT_AMT	decimal	0	14300000	13536.04	186809.80
38	L3DEP_SA_CREDIT_AMT	decimal	0	36700000	22330.96	336112.40
39	DEP_SA_DAYAVG_BAL	decimal	0	59706.82	2212.841	2430.461
40	L3_DEP_SA_DAYAVG_BAL	decimal	0	2241293	3077.981	15596.69
41	L6_DEP_SA_DAYAVG_BAL	decimal	0	1713113	3227.144	17538.19
42	L3_DEP_CARD_CUST_AMT	decimal	0	67600000	14384.64	372791.30
43	CHANNEL_OTHER_CREDIT_AMT	decimal	0	6000000	3839.155	89602.14
44	CHANNEL_OTHER_IN_MIN_AMT	decimal	0	85000	7.1320	481.6997
45	L6_CHANNEL_OTHER_AVG_CNT	decimal	0	236	1.0051	2.2349
46	L3_cr_dr_ratio	decimal	0	19.6	0.0480	0.3023

附表 3.4 中端客户建模指标基本统计描述

ID	Variable	Type	Min	Max	Mean	Std. Dev.
1	gender	varchar	0	2	1.5020	0.5198
2	LOAN_FLAG	varchar	0	1	0.0746	0.2627
3	DEP_TD_FLAG	varchar	0	1	0.3359	0.4723
4	DEP_SA_FLAG	varchar	0	1	0.9263	0.2613
5	CRED_FLAG	varchar	0	1	0.1821	0.3859
6	CUST_STAD_PLATINUM_FLAG	varchar	0	1	0.0667	0.2494
7	RELATED_REPAY_FLAG	varchar	0	1	0.0322	0.1764
8	CUST_GOLD_COMMON_FLAG	varchar	0	1	0.1283	0.3344
9	CUST_DOLLER_FLAG	varchar	0	1	0.2038	0.4028
10	CUST_SALARY_FINANCIAL_FLAG	varchar	0	1	0.0765	0.2658
11	FUND_FLAG	varchar	0	1	0.1283	0.3344
12	C_FUND_FLAG	varchar	0	2	0.1338	0.3563
13	LAST_OPEN_TENURE_DAYS	integer	0	88	3.0583	8.1682
14	CUST_PRODUCT_CNT	integer	0	15	2.6302	2.5478
15	CUST_ACCOUNT_OPEN_CNT	integer	0	1115	9.6983	20.4245
16	DEP_SA_OPEN_TENURE_DAYS	integer	0	6109	2183.207	1160.803
17	DEP_SA_LAST_TENURE_DAYS	integer	0	5936	1420.438	1122.99
18	DEP_SA_SUMACCOUNT_CNT	integer	0	523	2.2957	4.6843
19	DEP_SA_FGCR_ACCOUNT_CNT	integer	0	200	0.6697	1.8540
20	DEP_SA_ACCOUNT_CNT	integer	0	523	2.2957	4.6843
21	DEP_SA_DEBIT_CNT	integer	0	2341	5.0797	24.6577
22	DEP_SA_CREDIT_CNT	integer	0	479	4.6394	11.1448
23	L3DEP_SA_DEBIT_CNT	integer	0	4608	4.9098	34.3258
24	L3_DEP_CARD_CUST_CNT	integer	0	1318	6.1368	22.8273
25	CHANNEL_OTHER_CREDIT_CNT	integer	0	75	2.0891	2.9439
26	CUST_AUM	decimal	10000	199965.5	57276.47	48137.11
27	CUST_ASSET_AMT	decimal	10000	199965.5	57276.47	48137.11
28	L3_CUST_AVG_AUM	decimal	3334.603	4026593	60072.49	70799.11
29	L6_CUST_AVG_AUM	decimal	2052.745	7165060	70034.53	158879.80
30	L6_CUST_SAVING_AVGAMT	decimal	0	5478535	49952.95	94284.19
31	L3_CUST_ASSET_AVG_AMT	decimal	3334.603	4026593	60072.49	70799.11
32	L6_CUST_ASSET_AVG_AMT	decimal	2052.745	7165060	70034.53	158879.80
33	L6_CUST_DIFF_IN_AMT	decimal	0	37100000	25829.07	305690.00
34	DEP_SA_BAL	decimal	-1979934	3807333	740.2370	55823.64
35	L3_DEP_SA_BAL	decimal	0	1279332	19277.82	35417.86
36	L3_DEP_SA_NEW_AVG_BAL	decimal	-1663278	1269111	-239.1087	28940.15
37	L6_DEP_SA_NEW_AVG_BAL	decimal	-596224	634557.3	-248.5604	14752.80
38	DEP_SA_MOTH_MAX_IN_AMT	decimal	0	15000000	39730.64	247127.20
39	L3DEP_SA_MOTH_MAX_IN_AMT	decimal	0	45300000	80780.54	518397.20
40	DEP_SA_DAYAVG_BAL	decimal	0	2750187	19392.84	34406.23

ID	Variable	Type	Min	Max	Mean	Std. Dev.
41	L6_DEP_SA_DAYAVG_BAL	decimal	0	2507179	17101.70	41016.07
42	L3_DEP_CARD_CUST_AMT	decimal	0	94500000	73200.93	1193344
43	L3_CHANNEL_CTR_AVG_CNT	decimal	0	69.3333	0.6459	1.5625
44	CHANNEL_OTHER_IN_MAX_AMT	decimal	0	10000000	19757.09	146678.3
45	CHANNEL_OTHER_IN_MIN_AMT	decimal	0	50000	22.646	402.5562
46	L3_LG_TXN_AVG_CNT	decimal	0	259	1.1396	5.1499
47	cr_dr_ratio	decimal	0	48	0.1719	0.5754

附表 3.5 高端客户建模指标基本统计描述

ID	Variable	Type	Min	Max	Mean	Std. Dev.
1	gender	varchar	0	2	1.5687	0.4993
2	LOAN_FLAG	varchar	0	1	0.0696	0.2544
3	DEP_TD_FLAG	varchar	0	1	0.3996	0.4898
4	BOND_FLAG	varchar	0	1	0.0086	0.0924
5	DEP_SA_FLAG	varchar	0	1	0.9750	0.1563
6	CUST_STAD_PLATINUM_FLAG	varchar	0	1	0.1056	0.3073
7	RELATED_REPAY_FLAG	varchar	0	1	0.0396	0.1950
8	CUST_GOLD_COMMON_FLAG	varchar	0	1	0.1275	0.3335
9	CUST_DOLLER_FLAG	varchar	0	1	0.6298	0.4829
10	CUST_SALARY_FINANCIAL_FLAG	varchar	0	1	0.0463	0.2101
11	FUND_FLAG	varchar	0	1	0.2780	0.4481
12	C_FUND_FLAG	varchar	0	2	0.2981	0.4992
13	LAST_OPEN_TENURE_DAYS	integer	0	84	3.7520	9.0483
14	CUST_ACCOUNT_CNT	integer	0	143	8.5041	7.6900
15	CUST_PRODUCT_CNT	integer	0	15	5.1901	2.7141
16	CUST_ACCOUNT_OPEN_CNT	integer	1	646	28.1546	35.8639
17	DEP_SA_OPEN_TENURE_DAYS	integer	0	6111	2190.433	1124.629
18	DEP_SA_LAST_TENURE_DAYS	integer	0	5230	859.9391	867.3858
19	DEP_SA_FGCR_ACCOUNT_CNT	integer	0	42	2.0220	2.2527
20	DEP_SA_DEBIT_CNT	integer	0	1628	8.6964	29.8639
21	L3_DEP_CARD_CUST_CNT	integer	0	2419	8.0125	48.8141
22	L6_DEP_CARD_CUST_CNT	integer	0	6291	15.4413	99.5742
23	CHANNEL_OTHER_DEBIT_CNT	integer	0	74	1.7301	2.9761
24	CHANNEL_OTHER_CREDIT_CNT	integer	0	912	5.3103	11.9114
25	LG_TXN_CNT	integer	0	313	3.9541	9.2958
26	L6_CUST_AVG_AUM	decimal	41401.49	1.26E+08	1056397	2663322
27	L6_CUST_FINA_AVGAMT	decimal	0	50100000	388476.1	1081051
28	L6_CUST_ASSET_AVG_AMT	decimal	41401.49	1.26E+08	1056397	2663322
29	L3_CUST_DIFF_IN_AMT	decimal	0	80700000	179083.2	1422357
30	DEP_SA_BAL	decimal	-2E+07	20200000	3642.069	498043.2
31	DEP_SA_NEW_BAL	decimal	-2E+07	20200000	3733.744	498038.5

ID	Variable	Type	Min	Max	Mean	Std. Dev.
32	L3_DEP_SA_NEW_AVG_BAL	decimal	-7291335	6643246	1371.887	196635.6
33	L6_DEP_SA_NEW_AVG_BAL	decimal	-3712493	3778324	970.7179	103558.8
34	DEP_SA_DAY_MAX_IN_AMT	decimal	0	90800000	421674	1953356
35	DEP_SA_DAYAVG_BAL	decimal	0	22800000	110045	591267.2
36	L3_DEP_CARD_CUST_AMT	decimal	0	1.78E+08	277501.9	3028763
37	L3_CHANNEL_CTR_AVG_AMT	decimal	0	55600000	180158	1283150
38	L6_CHANNEL_CTR_AVG_CNT	decimal	0	336.3333	0.8234	4.2567
39	L6_CHANNEL_INTER_AVG_AMT	decimal	0	56900000	206731	1484841
40	CHANNEL_OTHER_IN_MIN_AMT	decimal	0	57636.37	47.1683	706.5108
41	L6_CHANNEL_CTR_DTAIN_MAXAMT	decimal	0	30500000	141408.5	812415.2
42	cr_dr_ratio	decimal	0	65.7299	0.5690	1.1409
43	L3_cr_dr_ratio	decimal	0	30	0.7618	0.7145

后 记

岁月不居，时光如流，三年的硕士生涯已近尾声，在我人生的这一重要阶段中，得到了许多老师、同学和朋友的大力帮助，在此我要向你们表达诚挚的感谢，谈谈自己的肺腑之言。

从研究生起始，我的导师便不余遗力的对自己进行监督和培养，从学术到实践都对自己关心备至。这篇文章虽作为我的毕业论文，但是论文的每一处都离不开了导师的辛苦培养以及自己背后的精细雕琢，三年来多次的参加课题项目、科研实践以及实习经历，更是锻炼了自己的学术科研能力与实践应用能力，最终才能完成毕业论文。

从选题之前，导师便说要创新，此时仍旧回荡耳边的话语至今记忆犹新“要么有研究问题和视角的创新，要么有学术观点上的创新，要么有方法改进上的创新，或者均有更好”，当然最后也选定了自己论文的题目，贴切的符合了导师的要求。开题之后，论文思路已经明显，但导师的提点总是让我眼前一亮，一个不经意的观念却能让自己发现自觉完美的论文还存在如此多的漏洞，一次小小的建议犹如醍醐灌顶般找到论文存在的逻辑问题，总之可以说我的论文不断充实和完善，是导师一次又一次的给与灵感与指导，其深厚的学术素养使我受益匪浅。预答辩期间，导师更是经常督促我的论文完成进度，我更是不会辜负导师的关心，力争不在格式、语言表达和逻辑结构上让老师失望，用自己所学提交“完美答卷”，是前期的积累使得我的论文不断充实和完善，并最终得以顺利完成。

在最后的定稿之后，郑重地向我的恩师致以最诚挚的谢意！更要感谢兰州财经大学统计学院三年来对我的培养以及在攻读硕士学位期间熟悉的老师和同学们，他们从各个方面给予我帮助和支持，使我在学业上更上一层楼，让我喜欢上了应用统计专业，成为我人生职业生涯的又一个起点！衷心地感谢大家！