基于数据挖掘的客户细分模型研究及应用

原慧琳，杜 杰，李延柯

（东北大学 信息科学与工程学院，辽宁 沈阳 110000）

摘 要：为企业更深入了解消费者的行为和偏好，帮助企业制定决策和发展客户关系，结合现有的客户细分方法，提出一种多指标客户细分模型。从宏观和微观角度，对传统指标进行优化，构建 RFMPA 多指标客户体系；采用熵值法客观赋权；采用因子分析降维；采用改进的 K-means 算法完成客户细分。利用大型连锁超市客户消费数据进行实证研究，对比数据实验结果表明，该模型能够更好解决客户细分问题，提高企业客户关系管理和决策质量。

关键词：聚类；客户细分；数据挖掘；多指标；RFMPA 模型

中图法分类号：TP391 **文献标识号**：A **文章编号**：1000-7024（2021）01-0057-08

doi：10.16208/j.issn1000-7024.2021.01.009

Research and application of customer segmentation model based on data mining

YUAN Hui-lin, DU Jie, LI Yan-ke

(College of Information Science and Engineering, Northeastern University, Shenyang 110000, China)

Abstract: To understand consumers' behaviors and preferences more deeply and help enterprises make decisions and develop customer relationships, a multi index customer segmentation model was proposed based on the existing customer segmentation methods. Through data analysis technology, from the macro and micro perspectives, the traditional indicators were updated and refined to build an RFMPA customer indicator system. Objective weighting was implemented using entropy method. Data dimension reduction was carried out using factor analysis. The improved K-means algorithm was used for customer segmentation. Using the customer consumption data of a large supermarket chain for empirical research, and comparing the data experimental results, the model can better solve the problem of customer segmentation, improve the quality of enterprise customer relationship management and decision-making.

Key words: clustering; customer segmentation; data mining; multi-indicator; RFMPA model

0 引 言

如今，零售行业的市场竞争日趋激烈，这给企业带来了巨大压力，迫使他们需要更有效了解客户需求，以获得或保持该行业的竞争优势。为了提高客户的忠诚度和满意度，提供个性化的服务和制定精准的营销策略对企业来说是至关重要的。在现代消费者的偏好和品味中，企业不可能完全满足每一位消费者。然而，大数据时代的到来，为企业提供了使用数据分析和挖掘技术的机会，通过这些海量数据对客户进行细分，从而提高企业决策质量[1]。

传统的细分模型虽然在客户分类方面表现良好，但忽略了客户购买行为的周期性和产品的购买力，而这两个方面体现了客户价值信息。此外，在经典的 RFM 模型中，定义的时间变量只考虑客户的最新交易行为，但在很多情况下，因为客户的消费行为表现出时间上的变化，这样的变量并不能准确地反映客户的重复购买或访问倾向[2]。为了弥补上述不足，本文旨在研究一套基于数据挖掘的多指标客户细分模型，提高细分模型的准确度。利用熵值法赋权值，构建新的指标矩阵，并用因子分析法进行新指标矩阵的降维，减小算法的时间复杂度。最后，利用改进的

收稿日期：2019-08-12；**修订日期**：2019-12-05

基金项目：东北大学-永辉超市产学研战略合作基金项目（7043902891801）

作者简介：原慧琳（1969-），女，河北秦皇岛人，博士后，教授，CCF 会员，研究方向为数据挖掘、机器学习；杜杰（1994-），男，山东淄博人，硕士研究生，研究方向为数据分析与挖掘；李延柯（1994-），男，辽宁海城人，硕士研究生，研究方向为智能优化算法。

E-mail: yhl@neuq.edu.cn

K-means 算法实现客户分类。

通过某零售商提供的过去几年的销售点（POS）数据，验证我们的多指标客户细分方法可以有效地识别客户群体，帮助企业提高决策质量和客户关系管理水平。

1 相关知识

1.1 客户细分

客户细分是指企业在特定的市场环境和运营模式下，按照客户的行为、属性、需求及偏好等变量进行划分，并为其提供满足需求的服务和产品的过程。关于客户细分的研究主要从以下 4 个方面展开，包括客户行为、人口统计方法、生活方式细分以及利益的细分方法。目前，基于客户行为的细分方法最为广泛，该方法以信息技术为基础，利用数据库中已有的客户行为数据完成客户细分。最常用的为 Hushes 提出的基于 RFM（Recency、Frequency 和 Monetary）模型的客户细分方法。例如，Dursun 和 Caber 利用 RFM 模型，对酒店客户关系管理系统中的客户消费行为信息进行价值细分[3]。Krishna 和 Ravi 利用 RFM 模型进行客户细分，帮助企业根据客户的需求定制产品和服务，提高客户体验和满意度[4]。Cho 等认为客户的重要性并不相同，因此提出了加权的 RFM 模型，从客户的消费数据中挖掘行为模式，以提高推荐的准确性，完成客户细分[5]。

其次，客户细分另一个重要的问题是指标体系的划分。根据客户细分变量将整个客户群划分为不同的小群体，由具有相似需求和特征的客户组成。例如：Park 等提出了一种用于多类别背景下客户细分的模型框架，以预测客户购买模式[6]。Kwac 等根据客户的用电数据进行生活方式的细分，并根据细分结果对哪些生活方式群体可以成为某些能源项目的良好候选提出建议[7]。Chen 等根据顾客在服务提供中的角色和行为来识别不同的细分市场，通过与客户建立密切联系，提高服务质量[8]。Han 等展示了分类变量属性在客户细分中的重要性[9]。

如今，客户细分不仅能够有效地识别关键客户群，而且帮助企业更深层次地了解客户行为和偏好。利用客户细分结果，帮助企业制定差异化的客户管理和营销策略，实现企业与客户的双赢。

1.2 数据挖掘

数据挖掘是指从大量的数据中发现隐藏于其中的信息的过程，如特征（Pattern）、趋势（Trend）及相关性（Relationship），也可以说是从数据中提取信息或知识。通过使用复杂的数据分析工具来突出大数据集下的信息结构，发现这些数据之间隐藏的潜在关系。对于客户消费数据来说，数据挖掘技术能够帮助企业更好维系客户关系，多属性和多维度地发现客户群体消费需求和行为模式的差异性，实现精准化的客户关系管理。数据挖掘技术主要有以下几个方面：聚类（Clustering）、分类（Classification）、回归分析（Regression analysis）、预测（Prediction）、关联规则（Association rules）[10-12]。

在数据挖掘技术中，客户关系挖掘常用的几种算法如下：聚类算法、分类算法和关联规则挖掘。聚类算法可以发现不同客户群体消费行为的差异性，帮助企业制定精准的营销策略。分类算法可以预测未来客户消费行为的趋势。关联规则挖掘可以找出客户与产品之间的关联性，指导企业进行交叉销售。其中，Hu 等利用关联规则，挖掘有价值的购买模式和客户群体[10]。Zhuang 等使用 3 种混合类型的数据聚类算法对客户进行细分，挖掘有用的客户相关信息来获得竞争优势[13]。Murray 等利用数据挖掘方法来识别历史嘈杂的传递数据中的行为模式，从而更好实现客户细分[14]。Tleis 等利用 K-means 聚类算法，实现有机食品市场的客户价值细分[15]。Peker 等通过 LRFM 模型聚类实现杂货零售行业的客户细分[16]。Lotko 等利用神经网络对维修服务行业的顾客忠诚度进行建模分析[17]。

数据挖掘技术已成为企业辅助决策的重要工具。有效的客户关系管理需要借助数据挖掘技术，实现客户信息的特征提取和价值分类。充分利用客户消费信息，能够提高客户忠诚度和关系管理的质量。同时，有效地分配资源，实现公司利润最大化，保持同行业的竞争力。因此，企业在客户关系管理中使用数据挖掘技术具有重要的意义。

2 模型构建

本节中，主要介绍了研究的模型和客户细分流程。主要包括以下几个步骤：①数据获取和预处理；②分析与建模；③模型评估与优化。其中，模型的创新性主要体现在“分析与建模”阶段，包括：构建 RFMPA 多指标细分体系、熵值法客观赋权值、因子分析降维、聚类实现客户细分。具体模型流程如图 1 所示。

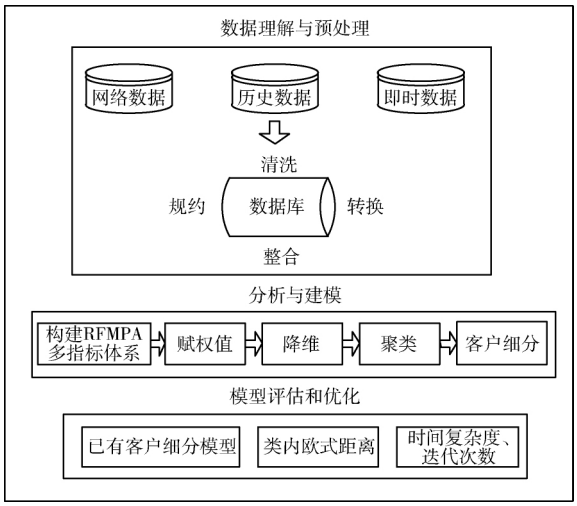

图 1 客户细分模型

2.1　数据获取和预处理

数据获取是数据挖掘工作的基础，是依据需求分析的结果提取、搜集数据，主要从网络数据与本地数据库中获得。但是，原始数据中存在大量异常数据，例如：数据缺失、异常值、不一致等，严重影响数据分析模型的效率，甚至会导致分析结果的偏差。所以，数据清洗变得尤其重要。数据清洗完成之后，接下来需要进行的是数据的转换、集成、规约等一系列操作，这就是数据获取和预处理。数据预处理一方面可以提高基础数据的质量，另一方面能够让数据更好地适应特定的数据挖掘模式，降低模型所花费的时间。

2.2　分析与建模

2.2.1　构建 RFMPA 指标体系

经典的 RFM 模型中，只考虑客户近期的交易情况，并不能完全刻画客户整体行为特征。结合数据的多维特性，我们对传统的客户细分指标进行更新和优化，主要体现在以下几个方面：

（1）将每一个维度划分为宏观和微观方面。在宏观方面可以反映出客户在整体消费中的情况，微观方面反映自身近期购买行为特征；

（2）增加了客户购买行为的周期性和数量。一方面能够准确反映客户交易行为，另一方面体现了客户的购买力。

首先，在 R（Recency）的选取上，经典的客户细分模型通常选取客户最近一次访问日期到观察期的时间间隔。在此基础上，我们将近度变量修改为客户 N 次访问日期到观察期之间的平均天数，可以观察出客户访问公司的程度，并提供有关重复购买倾向的信息，模型计算公式如下

$$Recency = \frac{1}{n}\sum_{i=1}^{n} date\_dis(t_{enddate}, t_{m-i}) \tag{1}$$

其中，$date\_dis(t_{enddate}, t_{m-i})$，表示观察期日期与客户来访日期之间的差值。$t_m$ 是客户最后一次访问。n 是客户总计访问的次数。当 $n=1$ 时，新定义的近度值变量等于传统的近度值，因此新特征变量包含了经典的变量特征。R1 为顾客消费平均近度值与所有客户平均近度值的比值，R2 为顾客一年内近度值与其自身历史近度值的比值。比值越高，说明客户消费时间距离观察期越近，客户的流失性越小。反之，客户的流失性越大。

F（Frequency）选取上，以顾客观察期内总的消费次数为分子，所有顾客的平均消费次数为分母，二者的比值记为 F1。宏观方面体现了客户在全部客户中的水平。微观方面，选取了近一年内总的消费次数和自身总的消费次数，二者的比值为 F2。目的是观察客户近期忠诚度的变化，如果近一年内总的消费次数与总的消费次数比值较大，说明客户的忠诚度处于上升期。

M（Monetary）选取上，统计客户在观察期内总消费金额，并计算所有客户的平均消费金额。M1 为客户消费总金额与全部客户平均消费金额的比值，M2 为客户近期内消费额与其历史总消费额的比值。通过消费金额的比值大小，可以观察出客户对企业的贡献度的高低。如果比值较大，说明客户购买力较大，企业应该将资源投入到这部分客户中去，提高客户满意度和客户价值。反之，客户购买力越小，企业应适当投放资源，并制定有效的营销策略，刺激客户消费。

在 P（Periodicity）的确定上，我们定义为客户访问间隔时间的标准差，它能够反映客户是否定期光顾商店，计算公式如下

$$Periodicity = stdev(VT_1, VT_2, \cdots, VT_n) \tag{2}$$

其中，n 表示客户访问间隔值的个数。VT 表示访问时间间隔，指客户连续两次访问之间经过的时间。P1 为客户购买产品的周期值与全部客户购买产品的平均周期值的比值。P2 为客户近期内购买产品周期性值与其历史总购买周期性值的比值。周期性表示客户访问是否倾向于定期进行。如果一个客户的周期性值较低，这意味着该客户访问或购买的时间间隔相对固定，可以被认为是有规律的。

A（Amount）为客户消费记录中购买商品数量的多少，A1 为客户购买产品数量与全部客户平均购买数量的比值。A2 为客户近期内购买产品数量与其历史总购买数量的比值。通过观察这一指标，目的是从客户的购买记录中发现客户消费的种类越多，那么对这类客户进行交叉销售可能性越高。在对商品购物篮分析之后，他们更倾向于购买种类较多的产品。企业可以根据客户的这种心理趋势来完成产品的交叉销售，提高产品销量。

构建 RFMPA 模型的指标体系见表 1。

表 1　客户细分指标体系

指标体系	细分类别
R(Recency)	R1，顾客消费平均近度值与所有顾客平均近度值的比值
	R2，顾客近期购买近度值与其历史近度值的比值
F(Frequency)	F1，顾客消费总次数与所有顾客平均消费次数的比值
	F2，顾客近期消费次数与其历史总消费次数的比值
M(Monetary)	M1，顾客消费总金额与所有顾客平均消费金额的比值
	M2，顾客近期内消费额与其历史总消费额的比值
P(Periodicity)	P1，顾客购买产品的周期性值与全部客户平均购买产品的周期性值的比值
	P2，顾客近期内购买产品周期性值与其历史总购买周期性值的比值
A(Amount)	A1，顾客购买产品数量与全部客户平均购买数量的比值
	A2，顾客近期内购买产品数量与其历史总购买数量的比值

2.2.2 确定权重

进一步研究，经典 RFM 模型在指标权重划分方面存在不同意见。Hughes 和 Arthur 认为 RFM 模型在权重划分方面是相同的，应该赋予相同的权重值。而 Stone 和 Jacobs 利用信用卡用户数据的实证分析表明，各个指标的权重并不相同，应赋予频度值最高，近度次之，花费金额最低。目前，关于客户细分指标权重的研究主要有以下两个方面：一是主观赋权法，包括层次分析法、特征值法等，主观评价法与决策者自身理解能力有关，人为因素的影响较大。二是客观赋权法，包括极差法、熵值法等，客观评价法重视数学理论的应用，从数据的离散程度和信息贡献度出发，不受决策者本身影响。

为了得到更加客观的客户细分结果，突出指标重要性，选用熵值法来计算细分指标的权重。按照各项指标观测值所提供信息的能力来确定权重值。熵值法的具体步骤如下：

(1) 建立数据矩阵

$$A=\begin{pmatrix} X_{11} & \cdots & X_{1j} \\ \vdots & \ddots & \vdots \\ X_{i1} & \cdots & X_{ij} \end{pmatrix}_{i\times j} \tag{3}$$

其中，X_{ij} 为第 i 个客户，第 j 个细分指标的数值。

(2) 数据标准化处理

其中，为避免计算熵值时对数的无意义，对数据进行了平移，正向指标

$$X'_{ij}=\frac{X_{ij}-min(X_{1j},\cdots,X_{nj})}{max(X_{1j},\cdots,X_{nj})-min(X_{1j},\cdots,X_{nj})}+1,$$
$$i=1,2,\cdots,n;j=1,2,\cdots,m \tag{4}$$

负向指标

$$X'_{ij}=\frac{max(X_{1j},\cdots,X_{nj})-X_{ij}}{max(X_{1j},\cdots,X_{nj})-min(X_{1j},\cdots,X_{nj})}+1,$$
$$i=1,2,\cdots,n;j=1,2,\cdots,m \tag{5}$$

(3) 计算第 i 个客户，第 j 个指标的比重

$$P_{ij}=\frac{X'_{ij}}{\sum_{i=1}^{n}X'_{ij}},(j=1,2,\cdots,m) \tag{6}$$

(4) 计算第 j 项指标的熵值

$$e_j=-k\sum_{i=1}^{n}P_{ij}\ln(P_{ij}) \tag{7}$$

其中，$k=1/\ln n$，ln 为自然对数，n 为客户数量，$e_j\geqslant 0$。

(5) 计算第 j 项指标的差异系数

$$g_j=1-e_j \tag{8}$$

对于第 j 项指标，指标值 X'_{ij} 的差异越大，对方案评价的作用越大，熵值就越小，g_j 值就越大。说明指标越重要。

(6) 计算各项指标的权值

$$W_j=\frac{g_j}{\sum_{j=1}^{m}g_j},j=1,2,\cdots,m \tag{9}$$

令 $W=\begin{pmatrix} W_1 & & \\ & \ddots & \\ & & W_m \end{pmatrix}$，将其与原先指标矩阵进行计算，即

$$A'=AW \tag{10}$$

2.2.3 数据降维

熵值法是按照各项指标的差异程度来确定权重值的大小，避免了主观因素带来的偏差，但熵值法并不能降低评价指标的维度，存在聚类时间复杂度较高的现象，所以我们引用了因子分析法对新指标矩阵进行数据降维。

因子分析模型：一般地设 $X=(x_1,x_2,\cdots,x_p)'$ 为可观测的随机变量，且有

$$X_i=\mu_i+a_{i1}f_1+a_{i2}f_2+\cdots+a_{im}f_m+e_i \tag{11}$$

其中，$f=(f_1,f_2,\cdots,f_m)'$ 为公共因子，$e=(e_1,e_2,\cdots,e_p)'$ 为特殊因子，f 和 e 均为不可直接观测的随机变量。$\mu=(\mu_1,\mu_2,\cdots,\mu_p)'$ 为总体 X 的均值。$A=(a_{ij})_{p*m}$ 为因子载荷矩阵。

通常先对 X 做标准化处理，使其均值为零，方差为 1，这样就有：

假定：

(1) f_i 的均数为 0，方差为 1；

(2) e_i 的均数为 0，方差为 δ_i；

(3) f_i 与 e_i 相互独立。

则称 X 为具有 m 个公共因子的因子模型。

如果满足 f_i 与 f_j 相互独立 $(i\neq j)$，则称该因子模型为正交因子模型。正交因子模型具有如下特性：

X 的方差可表示为

$$Var(x_i)=1=a_{i1}{}^2+a_{i2}{}^2+\cdots+a_{im}{}^2+\delta_i \tag{12}$$

设

$$h_i{}^2=a_{i1}{}^2+a_{i2}{}^2+\cdots+a_{im}{}^2 \tag{13}$$

则：

(1) $h_i{}^2$ 是 m 个公共因子对第 i 个变量的贡献，表示第 i 个共同度或共性方差；

(2) δ_i 为特殊方差，表示不能由公共因子解释的部分。

因子载荷是随机变量与公共因子的相关系数。

设

$$g_j{}^2=\sum_{i=1}^{p}a_{ij}{}^2,j=1,2,\cdots,m \tag{14}$$

称 $g_j{}^2$ 为公共因子 f_j 对 X 的“贡献”，是衡量公共因子重要性的一个指标。

因子分析步骤：

(1) 输入原始数据 X_{n*p}，计算样本均值和方差；

(2) 求样本相关系数矩阵 $R=(r_{ij})_{p*p}$；

(3) 求相关系数矩阵的特征根 $\lambda_i(\lambda_1,\lambda_2,\cdots,\lambda_p>0)$ 和相应的标准正交的特征向量；

(4) 确定公共因子数；

（5）计算公共因子的共性方差 h_i^2；

（6）对载荷矩阵进行旋转，以求能更好解释公共因子。

因子分析法是利用变量与变量之间的关系，用少数几个因子去表示多指标之间的相关性。Kaiser 度量标准见表 2。

表 2　因子分析度量标准

检测类别	值的范围	因子分析适合情况
KMO	0.90～1.00	Marvelous
	0.80～0.89	Meritorious
	0.70～0.79	Meddling/Middle
	0.60～0.69	Mediocre
	0.50～0.59	Miserable
	小于 0.5	Unacceptable
Bartlett's test	≤0.01	Acceptable

对新指标矩阵，我们根据 KMO 和 Bartlett's test 来确定变量之间是否适合进行因子分析，参照并通过累计方差贡献率和特征根来确定因子的数目，累计方差贡献率一般要不小于 85%，特征根要求大于 1。

2.2.4　聚类

接下来，需要对因子变量进行聚类，完成客户细分。其中常用的聚类算法为 K-means 算法，通过随机选取一组初始聚类中心，不断更新迭代，直到聚类结果不再变化[18]。但 K-means 算法中 K 值的确定是难以估计的，起初我们并不确定将数据集划分成多少个类别最合适，有些根据研究经验来确定 K 值。此外，聚类算法中初始中心点的选择对分类结果影响较大，如果初始值选取不好，可能无法得到预期的效果。所以，我们利用改进的 K-means 算法来弥补以上不足。

首先，根据 SSE（手肘法）确定最佳聚类数目 K，SSE 定义为每个簇的对象与其聚类中心之间距离的平方和。通常类别越多，SSE 就越小。一个合适的 K 值可以定义为 SSE 下降速度显著放缓的值。因为当 K 值小于真实聚类数时，由于 K 的增加会大幅提高每个簇的聚合程度，所以 SSE 的下降趋势很明显。而当 K 值达到真实聚类数目时，再增大 K 所得到的聚合程度会极速变小，SSE 的下降幅度也会骤减。所以说 SSE 和 K 的趋势图是一个手肘的形状，而肘部的位置就是对应的 K 值的真实聚类数[19]。

此外，当确定好聚类数目之后，在初始点的选择上，我们选取尽可能远的 K 个点，这个改进虽然简单直观，但却十分有效。具体算法如下描述：

（1）从输入的数据集中随机选取一个点，作为初始聚类中心点；

（2）对数据集中的每一个点 X，计算其与初始聚类中心点的距离 $D(x)$，并将其放到一个数组里边，然后距离相加得到 $Sum(D(x))$；

（3）选择下一个新的聚类中心点，选择原则是：$D(x)$ 较大的点，也就是距离初始中心点最远的点，被选取的机率较大。通过权重的方法来获取下一个初始种子点。步骤如下：

1）取一个可以落在 $Sum(D(x))$ 中的随机值 Random，计算方法为 $Sum(D(x))$ 与 0 到 1 之间的随机数相乘；

2）找出当前 Random 所在的区间，Random 等于 Random 减去 $D(x)$，直到其小于或等于 0，此时对应的点就是下一个初始种子点。如图 2 所示，Random 有更大的概率落在 $D(x_3)$ 中。

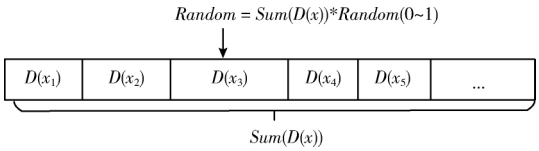

图 2　初始聚类中心点选取

（4）重复第（2）步和第（3）步，直到选出 K 个初始聚类中心点。

（5）根据选出的 K 个初始聚类中心点，运行标准的 K-means 算法。

另外，在距离计算方面，我们采用欧氏距离

$$D_{ij} = \| X_i - C_j \| = \sqrt{\sum_{u=1}^{n} | x_{iu} - c_{ju} |^2} \tag{15}$$

其中，X_i 为样本 i 所有指标形成的向量，C_j 是簇 j 的中心点对应这些指标的向量，n 是指标的个数。

2.3　模型评估

为了验证结果的有效性，我们与经典 RFM 指标对客户进行细分的结果对比。并验证选取初始中心点之后的聚类时间和迭代次数的优化。在聚类效果的评估中主要考虑的是类别的紧密程度，因此我们将每个客户点与其聚类中心点的类内平均欧氏距离作为标准

$$\bar{d} = \frac{\sum_{i=1}^{m} \sqrt{\sum_{u=1}^{n} | x_{iu} - c_{ju} |^2}}{m} \tag{16}$$

X_i 是样本 i 所有指标形成的向量，C_j 是簇 j 的中心点对应这些指标的向量，n 是指标的个数，m 为类内样本的数量。

3　模型应用

3.1　数据预处理

我们将某零售商提供的过去 3 年的 Pointofsales 数据（POS 数据）作为案例，数据集含有 3 万多条会员信息，约 38 万条消费记录。我们对原始数据集进行预处理工作，对存在多个属性信息缺失的情况予以删除，少部分缺失进行

插值补全。通过对数据的清洗和整合，最终有 31 099 条会员基本信息和 362 368 条消费信息被保留，约 94%的原始数据集。

3.2 分析与建模

首先，根据熵值法得到的每个指标的权重为 $W=$(0.12044223,0.13227003,0.00438084,0.26321809,0.00650823,0.16555703,0.00389175,0.14118024,0.00505721,0.15749435)。将得到的权重值按式(10)计算得到新的数据矩阵。

接下来，利用 KMO 和 Bartlett's 检验来确定新的数据矩阵之间是否适合进行因子分析。通过计算，我们得出的结果见表 3。

表 3 KMO 和 Bartlett's test

KMO Measure of Sampling Adequacy		0.857
Bartlett's Test of Sphericity	Approx. Chi-Square	4.258E4
	Sig.	0.000

可以看出 KMO=0.857，说明数据矩阵比较适合进行因子分析。Bartlett's test Sig 值小于 0.05，说明拒绝零假设，即相关矩阵不是单位矩阵，原矩阵之间有共同因素存在，适合进行因子分析。进一步，通过计算累计方差贡献率和特征根来确定因子的数目，见表 4。

表 4 总方差解释

成份	初始特征值			旋转平方和载入		
	合计	方差%	累积%	合计	方差%	累积%
1	5.74	57.42	57.42	4.25	42.56	42.56
2	1.59	15.96	73.39	2.52	25.26	67.82
3	1.36	13.60	87.00	1.91	19.18	87.00
4	0.59	5.923	92.92			
5	0.26	2.599	95.52			
6	0.21	2.108	97.63			
7	0.11	1.131	98.76			
8	0.05	0.54	99.31			
9	0.05	0.51	99.81			
10	0.02	0.18	100.0			

从表 4 可以看出，因子 1 的方差百分比为 57.42%，因子 2 的方差百分比为 15.96%，因子 3 的方差百分比为 13.605%，前 3 个因子累积贡献率为 87%。另外，观察特征值和旋转平方和载入数据，最终我们选取了 3 个因子。

聚类方面，为了弥补传统聚类算法的不足，我们首先根据 SSE 法来确定最佳聚类的数量，通过观察肘部的位置来确定 K 值。将降维后，选取 3 个公共因子的数据集作为输入，找出肘部位置，如图 3 所示。显然，肘部对应的 K 值为 5，所以针对这个数据集来说，最佳聚类数目应该选择5 类。

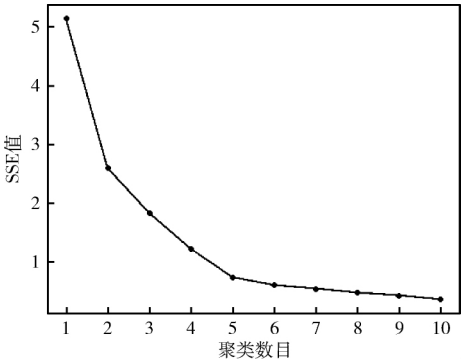

图 3 SSE 图

接下来，在初始聚类中心点的选择上，按照 2.2.4 节所描述，选取尽可能远的 5 个点作为初始聚类中心，结果见表 5。

表 5 初始聚类中心

	1	2	3	4	5
FAC1	−1.024 28	2.942 18	−0.547 40	−0.173 03	−0.501 72
FAC2	0.953 39	0.339 09	−1.679 33	0.916 02	0.957 67
FAC3	−1.555 57	0.516 87	−0.142 32	1.551 66	−0.192 59

最后，我们根据标准 K-means 算法，将客户分为 5 类，聚类信息见表 6。

表 6 多指标客户细分结果

类别	近度	频次	金额	周期	数量	样本数	$\bar{d}$
C1	2.1	2.7	9116.9	3.5	16.4	9458	0.84
C2	4.1	3.4	12 790.3	4.0	23.2	4706	0.69
C3	3.0	1.6	4832.4	2.2	8.6	6641	0.27
C4	4.8	7.7	36 356.1	4.5	63.0	2434	0.97
C5	2.6	1.1	5785.9	2.8	10.3	7860	0.57
总计						31 099	

依据聚类结果，我们将客户细分为 5 个等级，分别为：C1 中价值客户、C2 重要发展客户、C3 低价值客户、C4 高价值客户、C5 一般客户。

3.3 模型评估

对同一个数据集，根据经典客户细分指标对客户进行细分，并按照本文所述评估方法进行计算，聚类信息见表 7。同时监控新模型和经典模型的算法运行时间和迭代次数，以及聚类中心变动大小的变化。

通过实际案例发现，在找出初始聚类中心以后，聚类中心变动均值（5 个聚类中心点变化的平均值）从最初的

表 7　传统细分模型结果

类别	近度	频次	金额	样本量	$\bar{d}$
C1	2.19	1.69	7069.11	8312	0.77
C2	4.46	1.41	5406.69	12 217	0.56
C3	2.99	5.90	15 516.88	7105	0.85
C4	2.59	7.06	21 398.70	2631	1.30
C5	2.52	10.35	34 077.13	834	2.99
总计				31 099	

1.87 下降到 0.57，说明初始点的选取对聚类迭代有很大的影响。从图 4（横坐标为聚类迭代次数，纵坐标为聚类中心变动的均值大小）结果对比可以看出，标准的聚类算法迭代了 70 多次，而加入初始点以后迭代了 30 多次，聚类的迭代次数是原来的 1/2 左右，说明对初始聚类中心点的选取做了优化，简要来说就是使初始聚类中心点尽可能分散开来，这样可以有效减少迭代次数，加快运算速度。而且聚类所花费时间从 00：01.13 下降到 00：00.50，可以看出算法在改进之后迭代次数和聚类时间都得到了优化。

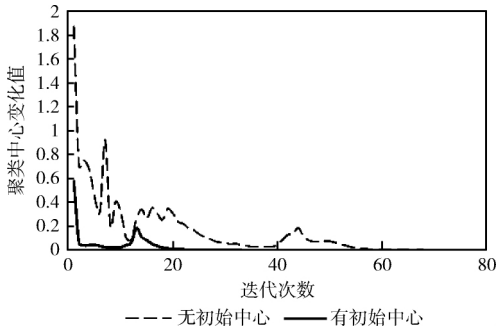

图 4　结果对比

我们将两种模型的细分结果进行对比，很容易发现，经典 RFM 模型细分结果中每个类别除了花费金额差异较大，其它特征差异较小。另外，可以看出各个类别的类内平均距离较大。而通过表 7 可以看出，利用多指标客户细分模型得到的细分结果，类与类之间的差异较大，类内差异较小，聚类效果更为紧凑。这表明该模型在聚类紧凑性和特征划分能力方面优于传统的客户细分方法，可以有效地帮助企业区分不同类型的客户群体，提高客户关系管理水平和决策质量。

4　客户细分策略

本文提出的多指标客户细分模型，根据细分结果，可以帮助企业决策者制定精准的营销策略，加强企业与客户之间的联系，从而带来更高的利润。在本节中，我们将提供基于客户细分的管理策略示例，目的是留住高价值客户，吸引一般客户，争取重要发展客户，从而提高企业利润和客户满意度。

中价值客户（C1），他们是企业比重最大的客户，占整体的 30%左右，消费水平是整体客户的平均水平。然而这个群体中，客户消费的平均近度值较低，说明客户购买产品的时间间隔较长，流失的可能性较高。企业应该关注这类客户的最新消息，采取一定的营销方法，降低客户流失的可能性。

重要发展客户（C2），他们是企业的潜在价值客户，客户人数占整体的 15.1%。虽然消费水平低于高价值客户，但整体来看属于企业的忠实客户，有很大的发展潜力。在营销活动中，企业应重视与这类客户的关系，制定适当的用户策略，刺激他们消费。另外，促进重要发展客户向高价值客户转变，实现企业长远稳定的收益。

低价值客户（C3）和一般客户（C5），这两类客户人数占了总人数的 50%左右。整体表现为购买数额小、频次低、时间间隔较远，购买行为具有很大的随意性。通常，商品促销和降价对这类客户有很大的吸引力。企业可以定期制定营销活动，促进他们向发展客户的转变。同时，企业应该适当减少这类客户的资源投入，转移到有价值的客户群体，从而达到企业资源的有效利用。

高价值客户（C4），他们的购买金额大，消费频次多，购买种类多，对企业的贡献最大，但他们所占的比例却最小，占整体客户的 7.8%。企业在进行客户关系管理时，应该重点关注这类客户。将企业资源优先投放到他们身上，并进行个性化管理和精准的营销策略，提高他们的满意度和忠诚度，延长这类客户的消费周期。

5　结束语

本文针对当前客户细分的背景，结合数据挖掘工具，提出了多指标客户细分模型。从微观和宏观角度考虑，将传统指标进行细化，并加入新的细分指标。通过熵值法为指标赋权。为了减少聚类的时间复杂度，利用因子分析进行数据降维。最后，利用改进的 K-means 聚类算法，在 K 值的确定和初始中心点的选取上进行优化，确定客户细分结果。对某零售商会员数据进行细分的实证研究结果表明，在聚类紧凑性和特征划分能力方面优于经典的客户细分方法，能够帮助企业提高客户关系管理水平和决策质量。

客户细分有助于公司的战略制定并提升竞争力。为了更好满足客户需求和偏好，企业必须认识到客户的差异性，从而制定精准的营销策略。基于客户细分问题的研究，未来的工作将围绕更加细致的客户分类，分析不同客户具有的各种用户特征。结合数据挖掘技术，辅助客户细分的决策与优化。我们将进一步对上述问题进行研究，期望获得更有理论意义和实际应用价值的成果。

参考文献：

[1] Kowalczyk M，Buxmann P. An ambidextrous perspective on

business intelligence and analytics support in decision processes: Insights from a multiple case study [J]. Decision Support Systems, 2015, 80 (10): 1-13.

[2] Sarantopoulos P, Theotokis A, Pramatari K, et al. Shopping missions: An analytical method for the identification of shopper need states [J]. Journal of Business Research, 2016, 69 (3): 1043-1052.

[3] Dursun A, Caber M. Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis [J]. Tourism Management Perspectives, 2016, 18 (3): 153-160.

[4] Krishna G J, Ravi V. Evolutionary computing applied to customer relationship management: A survey [J]. Engineering Applications of Artificial Intelligence, 2016, 56 (1): 30-59.

[5] Cho Y S, Moon S C. Weighted mining frequent itemsets using FP-tree based on RFM for personalized u-commerce recommendation system [J]. Lecture Notes in Electrical Engineering, 2014, 274 (1): 441-450.

[6] Park C H, Park Y H, Schweidel D A. A multi-category customer base analysis [J]. International Journal of Research in Marketing, 2014, 31 (3): 266-279.

[7] Kwac J, Flora J, Rajagopal R. Lifestyle segmentation based on energy consumption data [J]. IEEE Transactions on Smart Grid, 2018, 9 (4): 2409-2418.

[8] Chen S C, Raab C, Tanford S. Segmenting customers by participation: An innovative path to service excellence [J]. International Journal of Contemporary Hospitality Management, 2017, 29 (5): 1468-1485.

[9] Han S, Ye Y, Fu X. Category role aided market segmentation approach to convenience store chain category management [J]. Decision Support Systems, 2014, 57 (1): 296-308.

[10] Hu Y H, Yeh T W. Discovering valuable frequent patterns based on RFM analysis without customer identification information [J]. Knowledge-Based Systems, 2014, 61 (1): 76-88.

[11] Wong E, Wei Y. Customer online shopping experience data analytics integrated customer segmentation and customized services prediction model [J]. International Journal of Retail & Distribution Management, 2018, 46 (4): 406-420.

[12] Bejaei M, Wiseman K, Cheng K M. Developing logistic regression models using purchase attributes and demographics to predict the probability of purchases of regular and specialty eggs [J]. British Poultry Science, 2015, 56 (4): 425-435.

[13] Zhuang K, Wu S, Gao X N. Auto insurance business analytics approach for customer segmentation using multiple mixed-type data clustering algorithms [J]. Tehnicki Vjesnik-Technical Gazette, 2018, 25 (6): 1783-1791.

[14] Murray PW, Agard B, Barajas MA. Market segmentation through data mining: A method to extract behaviors from a noisy data set [J]. Computers & Industrial Engineering, 2017, 4 (17): 233-252.

[15] Tleis M, Callieris R. Segmenting the organic food market in Lebanon: An application of k-means cluster analysis [J]. British Food Journal, 2017, 119 (7): 1423-1441.

[16] Peker S, Kocyigit A, Eren P E. LRFMP model for customer segmentation in the grocery retail industry: A case study [J]. Marketing Intelligence & Planning, 2017, 35 (4): 544-559.

[17] Lotko A, Korneta P A, Lotko M A, et al. Using neural networks in modeling customer loyalty in passenger cars maintenance and repair services [J]. Applied Sciences, 2018, 8 (5): 713-729.

[18] Huerta-Muñoz D L, Ríos-Mercado R Z, Ruiz R. An iterated greedy heuristic for a market segmentation problem with multiple attributes [J]. European Journal of Operational Research, 2017, 261 (1): 75-87.

[19] You Z, Si Y W, Zhang D, et al. A decision-making framework for precision marketing [J]. Expert Systems with Applications, 2015, 42 (7): 3357-3367.