基于改进的 K-means 算法在共享交通行业客户细分中的应用

- **目的：得到用户画像当中的用户价值模型**

在众多客户关系管理的分析模式当中，目前识别客户价值最广泛的模型是通过三个指标（最近消费时间间隔（Recency）、消费频率（Frequency）和消费金额（Monetary）来进行客户细分，识别出高价值客户，简称 RFM 模型[1]。

在分类方面，现在普遍采用的是聚类分析方法。目前使用最广泛的聚类算法是 K-means 算法[2]。文献[3]提出的 K-means 算法效果受聚类数、初始聚类中心等因子的影响较大，研究表明上述影响因子与具体案例与主观经验有关。

- **计算模型：共享交通的 LRFMD 客户细分模型**

在原有的 RFM 模型上，选择客户在一定时间内累计的行驶距离 M 和客户在一定时间内享受的折扣系数的平均值 D 两个指标代替原有的消费金额 M。此外，注册会员时间的长短在一定时间程度上能够影响客户价值，所以在模型中增加客户关系长度 L，作为区分客户的另一指标。

LRFMD 客户细分模型：

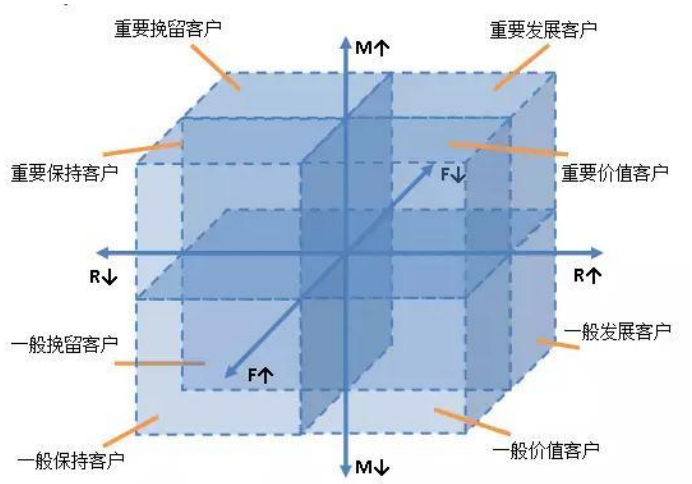
指标含义

模型	L	R	F	M	D
车辆 LRFMD 模型	会员注册时间距观测窗口结束的时间段	客户最近一次乘坐驾驶车辆距观测窗口结束的时间段	客户在观测窗口内驾驶车辆的次数	客户在观测窗口内累计的行驶里程	客户在观测窗口内驾驶车辆所享受的折扣系数的平均值

观测窗口：以过去某个时间点为结束时间，某一时间长度作为宽度，得到历史时间范围内的一个时间段。

针对的 LRFMD 模型，如果采用传统的 RFM 模型分析的的属性分箱方法，如图所示，它是依据属性平均值进行划分，其中大于平均值的表示为↑，小于平均值的表示为↓，该模型虽然也能识别出最有价值的客户，但是细分的客户群太多，提高了针对性营销的成本。因此本文采用聚类的方法识别客户价值。基于改进的 K-means 算法，通过对客户价值的 LRFMD 模型的五个指标进行聚类，

识别出最有价值客户。



RFM 模型分析

- **具体算法：K-means 算法细分用户**

1.数据处理

1.1 数据抽取

1.2 数据探索分析

探索分析是对数据进行缺失值分析与异常值分析，分析出数据的规律与异常值。查找每列属性观测值中空值个数、最大值、最小值的探索结果如下所示：

数据探索结果分析表

属性名称	空值记录数	最大值	最小值
User_id	0	47042	1939
Current_miles	0	13710	0
…	…	…	…
Cost	11	78.1	-29.4
Car_id	0	246	68

1.3 数据预处理

1.数据清洗

通过数据探索分析，发现数据中存在缺失值，异常值，Cost、Money 属性中存在值小于 0 的情况。由于原始数据量巨大，这类数据占比较小，对于问题影响不大，因此对其进行丢弃处理。具体处理方法如下。

- 丢弃缺失值
- 丢弃 Cost、Money 属性中值小于 0 的记录

2.属性规约

原始数据中属性太多（共 31 个属性），选择与 LRFMD 指标相关的 6 个属性，即 User_id，Start_time，Load_time，Cost，Money，bonus。

3.数据变换

数据变换是将数据转换成“适当的”格式，以适应挖掘任务及算法的需要。主要采用的数据变换方式为属性构造和数据标准化。（由于原始数据中并没有直接给出 LRFMD 模型的 5 个指标，需要通过原始数据提取这五个指标，如 L = Load_time － Start_time）

LRFMD 指标取值范围

属性名称	L	R	F	M	D
MAX	450	448	369	239968	111
MIN	1	1	1	0	0
AVG	257.39	168.54	14.57	9308.94	3.87

从表中的数据可以发现，5 个指标的取值范围数据差异较大，为了消除数量级数据带来的影响，需要对数据进行标准化处理。本文采用 Zscore 标准差标准化处理方式，处理结果部分数据一览，如下表所示。

标准化处理后的数据集

ZL	ZR	ZF	ZM	ZD
1.20882776	1.312803469	-0.185458018	-0.554793993	-0.526600821
0.956659445	1.497281062	-0.347784511	-0.554793993	-0.526600821
1.249094637	1.83109238	-0.510111005	-0.552708063	-0.322491976

0.872614763	1.463577819	-0.550692628	-0.554793993	-0.526600821
-1.581091175	-1.016106791	-0.550692628	-0.554793993	-0.526600821
1.033287101	1.625940291	-0.510111005	-0.554793993	-0.526600821
0.847914913	1.43860888	-0.510111005	-0.554793993	-0.526600821
1.230696342	-0.901523796	1.356643666	-0.091002535	2.085992401

2. K-means 算法及改进

K-means 聚类的目的是：把 n 个点（可以是样本的一次观察或一个实例）划分到 k 个聚类中，使得每个点都属于离他最近的均值（此即聚类中心）对应的聚类，以之作为聚类的标准[4][5]。这个问题在计算式是 NP 难，不过存在高效的启发式算法。一般情况下，都使用效率比较高的启发式算法[6]，它们能够快速收敛于一个局部最优解。已知观测集$(x_1, x_2, \cdots, x_n)$，其中每个观测都是一个 d 维实向量，K-means 聚类要把这 n 个观测划分到 k 个集合中$(k \leqq n)$，使得组内平方和(WCSS with-cluster sum of squares)最小。换句话说，它的目标是找到使得下式满足的聚类S_i:

$$\frac{arg\ min}{s}\sum_{i=1}^{k}\sum_{x\epsilon s_i}\|x-\mu_i\|^2$$

其中μ_i是S_i中所有点的均值。K-means 算法具体步骤如下。

输入：样本数据集 X 和聚类数 k

输出：k 个类

（1） 随机选择 k 个初始聚类中心；

（2） 逐个将数据集 X 中各点按最小距离原则分配给 k 个聚类中心的某一个；

（3） 重新计算每个类的聚类中心；

（4） 若新的聚类中心和原来的聚类中心相等或小于预设阈值，则计算结束，否则转步骤（2）。

改进初始聚类中心的选取方法

传统 K-means 聚类算法通过初始中心迭代得到最后的 k 个中心。这个初始中心可以随便选也可以随机选，也可以只取前 k 个样本作为初始中心。聚类

最后的结果与初始聚类中心的关系还是比较密切的，不同的初始中心可能会得到完全不同的结果。文献[7]提出了基于数据分段的思想来确定出事聚类中心。文献[8]提出了 K-means++基于最大概率的方式确定初始聚类中心。方法具体如下：

*（1） 从输入的数据点集合中随机选择 K 个点作为聚类中心，重复 5 次取样，得到 5*k 个样本点组成的集合 ，再聚类为 k 个初始中心点；*

（2） 对于数据集中的每一个点 x，计算它与最近聚类中心(指已选择的聚类中心)的距离 D(x)，并基于欧氏距离的最大概率准则选择新的聚类中心；

（3） 重复过程（2）直到找到 k 个聚类中心。

聚类数 K 值的选取

运用 K-means 算法时，需要预先给定聚类数 k,该算法是针对客户价值细分领域的，可以根据工程经验将 k 值取作 5。

聚类结果

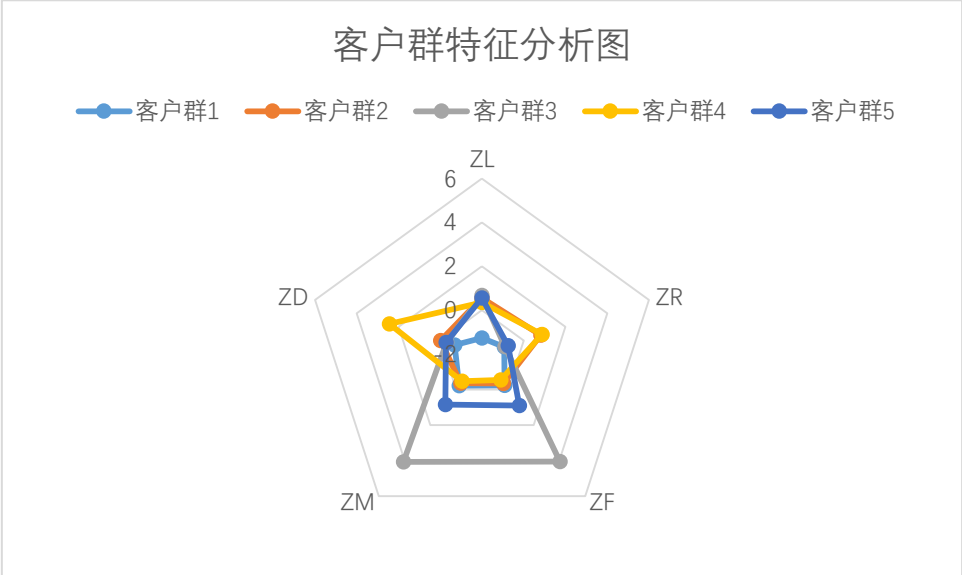
运用 K-means 算法对包含 L、R、F、M、D 各指标的标准化数据进行聚类，聚类结果如下所示。

客户分类情况

ZL	ZR	ZF	ZM	ZD	num	per
-1.26534934	-0.93586135	-0.24461487	-0.22321496	-0.68739483	1030	29.59%
0.56052897	0.81107607	-0.33850695	-0.35428804	-0.01618059	1415	40.65%
0.6649835	-0.90491723	4.04972263	4.07757598	-0.3143415	108	3.10%
0.3475817	0.87571874	-0.53503839	-0.46243334	2.44529302	376	10.80%
0.55734129	-0.75230912	0.89627725	0.84189442	-0.28001647	552	15.86%

3. 客户价值分析

针对聚类结果进行特征分析，如图所示。其中客户群 1 在 R 的属性上最小；客户群 2 在 R 属性上最大；客户群 3 在 L、F、M 属性上最大，R 属性上也较小；客户群 4 在 D、R 属性上最大。



客户群特征分析图

根据每种客户类型的特征，对各类客户群进行客户价值排名，其结果如下表所示。只可可以针对不同类型的客户群提供不同的产品与服务，例如提升重要发展客户的价值。

客户群价值排名

客户群	排名	排名含义
3（LRFM 值较大）	1	重要保持客户
1（R 属性最小）	2	重要发展客户
5	3	重要挽留客户
2	4	一般价值客户
4	5	低价值客户

参考文献

[1] 罗亮生，张文欣.基于常旅客数据库的航空公司客户细分方法研究[J].现代商业，2008（23）

[2] 贺玲, 吴玲达, 蔡益朝. 数据挖掘中的聚类算法综述[J]. 计算机应用研究, 2007, 24(1):10-13.

[3] 张静. 基于 K-means 聚类算法的客户细分研究[D]. 合肥工业大学, 2013.

[4] MacKay, David. Chapter 20. An Example Inference Task: Clustering. Information Theory, Inference and Learning Algorithms. Cambridge University Press.2003: 284–292. ISBN 0-521-64298-1. MR 2012999

[5] Since the square root is a monotone function, this also is the minimum Euclidean distance assignment.

[6] E.W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics. 1965, 21: 768–769.

[7] Liu C, Zeng L, Zhang J, et al. An optimized K-means clustering algorithm for CMP systems based on data set partition[J]. Journal of Computational Information Systems, 2015, 11(13):4727-4738.

[8] Bahmani B, Moseley B, Vattani A, et al. Scalable K-means++[J]. Proceedings of the Vldb Endowment, 2012, 5(7):622-633.